Statewide Data Strategy Report

Final Version 1.0

July 15, 2009

Office of the Chief Information Officer for the State of California

Revision History

REVISION HISTORY			
REVISION/WORKSITE #	**DATE OF RELEASE**	**OWNER**	**SUMMARY OF CHANGES**
DSR_DV1	04/03/2009	David Marx Linda Wells Venky Madireddi	Initial Draft Release
Final Version 1.0	7/15/2009	David Marx	OCIO Final

Approvals

NAME	ROLE	DATE
Dale Alvarez, Lee Mosbruker, Michael Byrne	OCIO Review / Draft updates	04/15/2009
Dale Alvarez	OCIO Review	07/15/2009

Table of Contents

Table of Contents (Charts)

Table of Contents (Tables)

Table of Contents (Figures)

1. INTRODUCTION

California has adopted six strategic concepts to focus the future direction of Information Technology (IT) in the State. The six strategic concepts are supportive of public priorities, statewide policy, and integrated IT initiatives. The strategic concepts anticipate a future in which Californians and their government utilize rich multi-media information – information that is secure yet widely available and easily accessible. The six strategic concepts are:

1. IT as reliable as electricity
2. Fulfilling technology's potential to transform lives
3. Self-governance in the digital age
4. Information as an asset
5. Economic and sustainable
6. Facilitating collaboration that breeds better solutions

It is vital that California's IT vision is aligned with these concepts. This alignment must begin at the strategy level, and be incorporated into the various identified directives. The California Statewide Data Strategy supports these six strategic concepts.

1.1 Executive Summary

Since the inception of the Office of Information Technology in 1983, the State has focused on increasing the use of interoperable data and information systems to enhance services, improve informed decision making, and reduce the cost of government operations. In 2005, the Office of Chief Information Officer (OCIO) for the State of California established a comprehensive Enterprise Architecture (EA) program for the State's Information Technology (IT). Since its inception, the EA program has developed a roadmap, published position papers, and launched several initiatives. One such initiative is directed towards defining a Statewide Data Strategy. This document presents the Statewide Data Strategy and describes the architecture and the implementation plan for a common and shared data environment that is consistent with the EA roadmap.

The quality of data impacts the quality of decisions made by the various agencies in the State of California. These decisions, in turn, directly affect the quality of life of the State's constituents. Moving the State of California to an enhanced data management future based on shared integrated data will improve the speed and quality of decision making and delivery of services to the State's constituents while reducing the cost associated with non-integrated systems.

The goals for the data strategy are:

- Define a data sharing environment to provide a single, accurate, and consistent source of data for the legislature, state agencies, and local governments for the services provided to the public.
- Define common standards for database management and integration with a view toward consolidation of data and software reuse.

- Define a framework that facilitates interfaces between state agencies and trading partners such as federal agencies, commercial entities, and local government.
- Identify the organizational changes needed within the state to institutionalize the aforementioned goals.
- Define a plan and approach to accomplish the next level of work needed to implement the data strategy.
- Support of the agencies' current initiatives with minimal disruptions. The data strategy was developed to ***support, not rewrite,*** the agencies' initiatives.

The strategy is very comprehensive and was developed based on input from various sources, including:

- Information from agency surveys
- Interviews with agencies
- Interviews with the Office of Information Security and Privacy Protection
- Federal Enterprise Architecture Data Reference Model (FEA DRM)[i]
- Industry best practices

The agency survey responses provided an understanding of the current data sharing environment. The goal was to determine what is working/not working related to delivery of accurate and consistent data for the Legislature, state agencies, and local governments for the services provided to the public. The OCIO identified eight agencies to participate in the survey. These agencies, taken collectively are believed to be representative of the statewide IT environment. The agencies selected were:

- Business, Transportation and Housing Agency (BT&H)
- California Department of Corrections and Rehabilitation (CDCR)
- California Department of Food and Agriculture (CDFA)
- California Environmental Protection Agency (CalEPA)
- California Health and Human Services Agency (CHHS)
- California Natural Resources Agency (CNRA)
- Labor and Workforce Development Agency (LWDA)
- State and Consumer Services Agency (SCSA)

The findings from the survey concluded that the current data sharing and data integration environment is not conducive to facilitate the EA vision. This is based on several findings; few adopted consistent technology standards, too many manual processing steps, little interface reuse, and limited coordination between the agencies. The findings of the survey are covered in detail in Section 2 of this document.

Another consideration for the data strategy was the FEA DRM. DRM is a business-driven, functional model for classifying data and information and defines how it supports the business of government. DRM provides a common, consistent way of categorizing and describing data to facilitate data sharing and integration, and describes the interactions and exchanges necessary between state, local, federal government agencies, and various customers, constituencies, and business partners.

The recommended strategy is based on the creation of a secure shared network. All state departments would have access to this network. At the core of the network data environment is *Shareable Data*. Shareable Data is defined to be data that is generated by one or more Lines of Business (LoB) and is accessible by authorized users statewide. Data assets such as system files, databases, documents, official electronic records, images, audio files, web content, and *Geographic Information System (GIS)* data[1] are to be treated as Shareable Data. The details of the LoBs and their sub-functions are published by OCIO and are available in California Business Reference Model (CalBRM). Shareable Data could cut across existing organizational boundaries similar to LoBs.

Implementation of a standards-based platform referred to as California Data Services (CDS) will provide the access to the *Sharable Data.* Shareable Data in CDS is cleansed and enriched in order to make it usable by different systems, applications, or users, irrespective of which department originally created it.

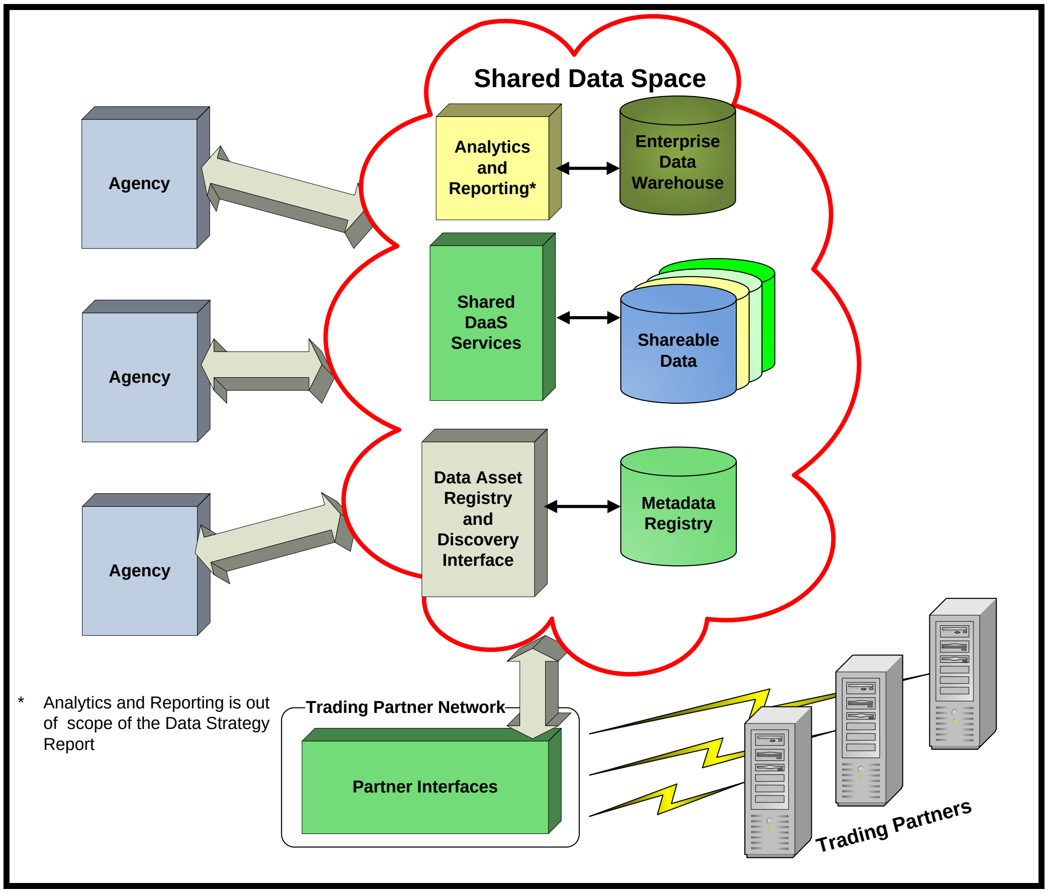
To enable integration with diverse technologies that are used statewide, the data strategy emphasizes the use of a *Data-as-a-Service (DaaS)* foundation as the basis of data sharing. The concept of DaaS evolved with the emergence of service-oriented architecture (SOA). SOA includes standardized processes for accessing data, and is independent of the actual platform on which the data resides. A subset of SOA is Service Oriented Integration (SOI) which facilitates integration between two computing entities using only service interactions. SOI supports web services and is a key ingredient in this strategy.

With SOI, any business process can access data using DaaS. A service oriented approach is suggested as the most straight-forward access methodology. Using a service oriented approach is not mandated since existing interfaces such as Java Database Connectivity (JDBC) or Open Database Connectivity (ODBC) would be supported to provide access. Since multiple copies of data exist throughout the agencies, a single, secure, validated, cleaned set of data will need to be maintained. Data ownership, update rules, security requirements, and data description will be defined by a data governance body. The implementation of these business requirements will be managed by a change management body. This 'single source of truth' data environment will be maintained externally to the agencies.

To facilitate discovery of the State's data assets, a metadata repository should be provided as a service. Services exposing data resources are to be made available for access to the network. Additionally, formal agreements to facilitate data sharing among agencies and departments through memorandums of understanding, or other appropriate agreements, will have to be established. Formal agreements to facilitate data sharing with trading partners will also have to be established. Agreements for sharing data between internal agencies are discussed in Section 5.6.7.5 and trading partners are discussed in Section 3.4.5.

[1] GIS data is also referred to as geospatial information.

Figure 1-1 is a conceptual diagram that illustrates the data environment and the components of the strategy. The architecture is discussed in detail in Section 3.



* Analytics and Reporting is out of scope of the Data Strategy Report

Figure 1-1 - Shared Data Space

The “cloud” in the diagram represents the shared secure network. *Shareable Data* represents a repository that holds data assets that are to be made “shareable” between agencies. *Metadata Registry* represents a data store that contains description of all data assets that are in Shareable Data. The group of services that enable service oriented access and update of data assets in Shareable Data is represented by *Shared DaaS Services*. The *Data Asset Registry and Discovery Interface* represents the interface to register and search data assets in Shareable Data. *Enterprise Data Warehouse* represents a consolidated warehouse to enable report generation for decision making by Legislature and state agencies. The *Partner Interfaces* in the Trading Partner Network would facilitate state agencies interface with trading partners such as federal agencies.

The recommended strategy for the State of California is to implement a secure shared network that facilitates data sharing, data integration, and warehouse consolidation. To facilitate discovery of the State’s data assets a metadata repository will be provided as a service. Since the agency survey conducted revealed that multiple copies of similar data exist and are being maintained within and across agencies, an effective strategy is predicated on *harmonization* of Shareable Data. Harmonization is the act of consolidating data from different sources according to the business rules that are established to enable a single, secure, validated, cleaned set of data. Harmonization rules, data ownership, security requirements, and data definition will be defined by a data governance body. The implementation of business requirements will be managed by a change management body.

There are nine initiatives that should be undertaken to implement the data strategy. These are:

1. Security Architecture
2. Infrastructure Configuration
3. Master Data Repository
4. Enterprise Content Management
5. Enterprise Service Bus
6. Develop Web Services
7. Metadata Repository
8. Trading Partner Framework
9. Data Warehouse Consolidation

A more complete description and specific recommended approach to be followed for each initiative is contained in the report, along with a high-level schedule and resource requirements. It is critical for the State to undertake these initiatives to implement the data sharing environment and to broaden the agency participation in the process. The detailed work plans for these initiatives, along with the development plan for the first set of services targeted towards Geospatial data are covered in Section 6.

The six strategic concepts that California has adopted are designed to support a future in which Californians and their government have access to secure, accurate, comprehensive, and rich multi-media information. These six strategic concepts and how the report is aligned with them are outlined below:

1. **IT as reliable as electricity** – The data strategy specifies a highly available infrastructure to provide reliable access to data. Poor quality data is a major contributor to poor IT reliability. The data strategy addresses data quality through centralizing and harmonizing the data across the agencies.
2. **Fulfilling technology's potential to transform lives** – Validated, high quality, and readily available data will be provided for critical State business decisions. One of the State's biggest IT challenges is providing consistent high quality data for the Legislature, decision makers, and constituents. The State's IT systems exist to automate business processes and to manage data. The data strategy identifies an approach that provides a high quality source of data to support the state government with the information needed for critical business decisions.
3. **Self-governance in the digital age** – High quality data will be made available for the constituents within the State, facilitating digital age decision making. Since the IT ecosystem within the State of California is so diverse, maintaining a single source of up-to-date data is difficult. The data strategy addresses this issue from both technical and business perspectives.

4. **Information as an asset** – Data within the State will be cataloged and made discoverable through an asset registry, so that information can be found and used for key decision making. Each agency manages their services using agency data. However, the agency data is really owned by the State and is a shareable asset, providing interagency agreements are in place. Having the ability to analyze information across all State agencies is currently a challenge, but the data strategy addresses this by providing a means of cleaning, accumulating, and sharing data in a consistent manner.
5. **Economic and sustainable** – The data strategy outlines economies of scale in shared infrastructure across agencies in both hardware and software. While addressing poor data quality can be expensive, making poor business decisions because of poor data quality can be devastating. The data strategy addresses this challenge and provides an efficient means of sharing data through standard services using mainstream interfacing techniques.
6. **Facilitating collaboration that breeds better solutions** – The strategy facilitates data sharing. Sharing data fosters collaboration between agencies and departments. In addition to data sharing, the data strategy addresses cross-agency data analysis. Since data is readily available across agency boundaries and agency purview, difficult questions can be analyzed that were simply not possible within a single stove-piped agency data source.

Special effort was made to provide an approach that incorporates risk management for the State. Given the budgetary constraints and the complex nature of the business of government, special focus has been given to developing a strategy that can be incrementally implemented and adheres to industry standards and best practices.

The data sharing strategy covered in the report:

- Responds to the findings from the agencies used as sampling sources for the project
- Defines an architecture, technology, and high level design for a secure data sharing strategy
- Defines how to leverage the FEA DRM principles within the strategy
- Identifies the standards and business concepts that are to be used in the strategy
- Identifies the approach necessary to implement the strategy
- Identifies the business and organizational changes needed to implement the strategy as well as the formal agreements that need to be in place
- Identifies the work plan to perform the work
- Identifies the risks and issues that will most likely be encountered when implementing the strategy

This report covers both the technological and business considerations necessary to provide an enterprise data sharing solution for the State.

1.2 Document Overview

The document is organized into eight major sections. They are:

Section 1 – Introduction
This section contains a brief summary of the findings and the data strategy itself. It also provides a high level description of what is contained in each of the data strategy report sections.

Section 2 – Existing Data Sharing Environment Review

This section provides a review of the existing statewide data sharing environment. It will identify the currently implemented data sharing solutions by agencies, and how they are working or not working towards delivering accurate and consistent data for the Legislature and other key decision makers.

Section 3 – Architecture and Design

This section provides an overarching architecture and functional design towards providing a single, accurate, and consistent source of data for the Legislature, State agencies, and local governments. It will also include an approach for consolidation of legacy data warehouses and data marts that exist statewide.

Section 4 – Strategy

This section provides an overall strategy towards achieving the implementation of "Architecture and Design" described in Section 3. The approach, where to start, and the initiatives that make the strategy a reality are identified.

Section 5 – Organizational Changes

This section offers recommendations on organizational changes necessary for achieving statewide data sharing, consolidating legacy data, managing change management, and improving governance.

Section 6 – Work Plan

This section provides a high level work plan for each of the initiatives detailed in Section 5. It includes manpower, timetable, and cost estimates. It also details the hardware and software resources that are required to complete the data strategy initiative as well as the dependencies between the other IT initiatives underway at the State.

Section 7 – Risk/Issues

This section addresses the immediate risks and issues that exist within the State, which can potentially derail the progress toward a secure statewide data strategy for California.

Section 8 – Definition of Standards

This section covers definition of standards for database management systems, internal framework for a Service Oriented Integration, and data interfaces with trading partners. It also offers recommended processes, guidelines, and policies to enable maximizing reuse of software and data.

This page intentionally left blank

2. Existing Data Sharing Environment Review

This section will provide a review of the existing statewide data sharing environment. It describes currently implemented data sharing solutions, and how they are working or not working towards delivering accurate and consistent data for the Legislature and other key decision makers. An analysis of the information obtained from the agencies with respect to existing data sharing solutions is discussed and recommendations made.

2.1 Overview

In order to assess the current data sharing environment, the project sampled the data structure and sharing arrangements at several State agencies. Eight State agencies provided information on over 800 applications that are used to run their business. Additional information was requested from the agencies on how data is shared between the applications and externally to other agencies and the federal government. Specific emphasis of the data sharing practices was targeted as follows:

- Data shared outside of the State
- Data shared between agencies
- Data shared between departments
- Data shared within a department

This section provides an analysis of the information provided by the agencies, and describes how data is shared and the technology used to share the data.

2.2 General Assessment

A questionnaire was conducted to start the assessment. The questionnaire was used to gain an understanding of the types of systems, types of data that exist, and how the data is used across the enterprise of the State of California. This section provides a general analysis of the information gathered from the agencies.

The questionnaire covers physical data sharing interfaces[2], and Table 2-1 – *Study Interface Count* represents the number of interfaces reported by each of the polled agencies.

[2] The interfaces that were requested were backend data sharing interfaces that moved data from one system to the other and were not user interfaces (i.e., screens).

Agency	Number of Interfaces Reported
BTH	536
SCSA	318
CHHS	252
LWDA	100
CDCR	87
CNRA	13
CalEPA	12
CDFA	8

Table 2-1 – Study Interface Count

The questionnaire allowed the agencies to provide only the information on the higher priority applications, so some interfaces were excluded from the response. The priorities of the applications were defined by the agencies themselves and were defined subjectively based on the business value of the application.

2.3 Observations

Based on the questionnaire results and follow-up interviews, it appears that all agencies are very competent at supporting their applications to ensure business continuity. The participating agencies use a wide variety of technology for applications and data interfaces (see Appendix C for complete details). While most of the data interfaces work well, it should be noted that most of the interfaces are of point to point design, with unique and rigid features. Very little Service Oriented Integration is in place, so we suspect the support costs are higher than they need to be, due to the rigid design of the interfaces.

Regarding improvements in data sharing, we observed the following from discussions with the agencies and the information provided to us:

- Coordination between departments and agencies is very limited
- Most data is still transferred via batch file based interfaces
- Interfaces requiring manual intervention are still prevalent
- Standardization is limited
- Cross-agency data sharing is limited

2.3.1 Limited Coordination

Little coordination exists between departments within an agency for interface development. For example, departments needing an interface to the federal government typically develop their own without coordinating or researching existing solutions that have been used in other departments. Even when there is research conducted to identify an existing interface, departments tend to ‘clone and modify’ the interface to their specific needs, rather than coordinate or partner with another department. Coordination on upgrades to these interfaces appears to be fragmented. For example, one department has 47 occurrences of a single set of similar interfaces to the federal government, all of which were 'cloned and modified'. Although this "clone and modify" approach may have evolved as the most expeditious approach, the long term maintenance costs are much higher than for a well documented shared interface.

2.3.2 File Transfer vs. Real Time

A vast majority of the interfaces reported by the agencies in the questionnaire involved flat file transfer of information and were batch interfaces. Transferring flat files has been used for years and provides a good means for moving data. However, newer techniques exist that provide a more robust and flexible means of transferring data. In most instances, data is transferred via a batch file transfer. Using the survey information, we determined the following composite data transfer profile:

Category	Count
Flat File Transfer	862
Real-time and pseudo real-time	236
Unknown	228

The comparison is not simply comparing a near real-time interface to a nightly batch interface as interface delays can be cumulative. For example, it takes one agency as much as 45 days to completely synchronize all data related to a *person*. Asynchronous web service interface with an eXtensible Markup Language (XML) payload is probably one of the more flexible interface approaches used today, and can be used to replace older file transfer techniques.

2.3.3 Manual Intervention

Manual intervention was common in all eight agencies. There were examples of manual interfaces required to kick off the process, reformat the data, or initiate the translation in some way. These manual steps are costly, and automating these data transfers should be considered wherever possible. The data also revealed that many agencies and departments are still extracting data to Microsoft Excel or Microsoft Access to do their data analysis. This approach also requires a significant amount of manual intervention. These manual processes are prone to errors and, although expedient, cost the State more in the long run. IT has been improving the efficiency of business through process automation for years. While manual intervention was a common theme, it is worth noting that within most agencies the data indicated a trend toward moving to fully automated interfaces.

2.3.4 Standardization

The State of California is struggling with the issue of standardization. This is largely due to the State's wide variety of services offered and vast technological landscape. If the State follows the lead of private firms, they will cut costs by limiting the technology used and supported. Several benefits can be realized by adopting this limited technology support technique. The top two benefits expected are efficiencies in IT support processes and cost savings through bargaining of licensing terms with IT vendors.

Although 'best in breed' solutions may be desirable from a functionality perspective, with the State's multifaceted business environment, taking this approach can introduce a very complex technical landscape. Where possible, with very few exceptions, a handful of technologies and vendors should be adopted that meet most of the State's IT needs.

Older technology that is outside of the accepted technology 'core list' for the State should have a replacement plan. A common issue with using older technology is that when support costs rise as qualified support personnel become scarce, it is also difficult, if not impossible, to easily replace an application that has evolved over many years with the same functionality. Many organizations find themselves 'stuck' with the application with spiraling IT support costs. To avoid this, every application should have a life expectancy and a replacement plan identified once it goes into service. With a plan in place, decisions to upgrade, add functionality, or extend an application are made in light of the bigger picture - the total cost of ownership of the application.

2.3.5 Cross-Agency Interaction

There is little incentive at an agency level for inter-agency collaboration. Most of the agencies evolved to support only their charter, which is the reason why most data is not shared. The bottom line is that data sharing was not designed into their original architecture - it was an afterthought. With this noted, it is no surprise that inter-agency interaction is minimal. Therefore, it is difficult to analyze cross-agency data relationships. For example, it is difficult to provide an accurate and comprehensive view of a constituent, in terms of their personal information, home addresses, services received, licenses granted, and organizational affiliations. Even if this analysis could occur, these relationships within the data are hard to maintain since there is little coordination occurring between agencies.

This challenge has been highlighted when trying to match up data from two different agencies, or possibly even from two departments in the same agency. For example, one department receives information regarding services offered by hospitals and another department receives information regarding services received by constituents at each hospital. Although each of the systems has high quality information, matching this information across both systems is extremely difficult due to how each stores its data. This makes it challenging for the State to see the 'big' picture.

2.4 Summary

Many organizations are limiting the technology within the IT infrastructure as well as limiting their interfacing solutions. By limiting the technology, support costs are driven down while improving the overall support and licensing costs are lowered through improved licensing terms with the vendor. Typical organizations will have at least a primary support person and a backup support person on any particular technology. Using more technologies requires more support staff and/or a support staff of generalists. Experts supporting a handful of technologies will provide timelier, higher quality services for lower cost than generalists supporting many technologies. The trend in the marketplace is to limit technology use to a few key technologies that can get the job done in most all instances. Exceptions will occur, but should be kept to a minimum. In an article featured in Network World, James Kobielus wrote “Fewer software licenses and servers translate into cost savings in capital and operating budgets. Fewer redundant software components translate into less need for redundant programming groups”. He goes on to quote Gartner Inc. “Application consolidation onto fewer platforms reduces software life-cycle costs, which can be six times greater than license costs"[ii].

The cost benefits are further substantiated in the following extract from a press release issued by Gartner Inc., "Organizations that have implemented substantial data integration architectures can save more than $500,000 annually by rationalizing tools in the short term and adopting a shared-services model in the longer term. Deployment of multiple and functionally overlapping data integration tools creates excessive cost in terms of software licensing, maintenance, and skills of up to $250,000 per tool annually.[iii]"

The following is an excerpt from a case study available at The Computerworld Honors Program, "In recent years, HUD has made substantial progress in migrating its once aging, 'stove-piped' infrastructure to a more efficient, shared services environment. By focusing its vision on evolving to a common enterprise architecture and, ultimately, a Service Oriented Architecture (SOA), HUD is leading the way to a cost-effective, shared services business-aligned IT environment that enhances mission effectiveness"[iv].

Another trend in the marketplace is toward use of more flexible asynchronous interfacing techniques to facilitate *Data as a Service* (DaaS). For example, using a web service leveraging an Open Source standard like XML should be considered for interfaces that move small to medium amounts of data. XML documents can be extended while maintaining some degree of backward compatibility.

For substantial updates or interfaces that move large amounts of data, Extraction and Transformation and Load (ETL) technology is a good approach for the interface design. ETL tools are well documented, well supported, and provide excellent data transformation and efficient data transfer capabilities.

This page intentionally left blank

3. Design and Architecture

This section describes the overarching architecture and functional design that can provide a single, accurate and consistent source of data for the Legislature, state agencies and local governments. It also describes an approach for consolidation of legacy data warehouses and data marts that exist statewide and a solution to integration that promotes reuse. This improved data sharing is provided in the form of Service Oriented Integration, a subset of SOA, and standard web services.

3.1 Overview

The State government within California is complex, and supports millions of constituents. The State government must also be nimble, as legislative changes can give rise to rapid changes with how government is conducted. To support these challenges, the State has adopted a policy of autonomy for each of its agencies. While this is a good approach, it does tend to lead to single-purpose, stove-piped solutions. As the pressure to be more efficient increases, so does the pressure to improve collaboration among the agencies and the constituents. This vision of a nimble California government that provides more effective constituent services with improved collaboration and efficiency is based on the implementation of a standards-based platform that we refer to as California Data Services (CDS).

The CDS design and architecture described in this strategy focuses on increasing interaction among the agencies and improving data quality and availability while providing a flexible solution to support legislative changes as they occur.

The following sections describe:

- The design principles
- The design
- The architecture

3.2 Design Principles

To realize the vision, two primary objectives must be emphasized: (1) increasing the data that is available along California Business Reference Model's (BRM[3]) Communities of Interest (COIs) / Lines of Business (LoBs) and (2) ensuring that data is usable by both anticipated and unanticipated users and applications. In order to achieve the objectives, the CDS design incorporates the following success principles:

- **Visibility -** Increase visibility of shareable data assets to users and applications. Both should be able to discover the existence of data assets through registries. All shareable data assets (structured and unstructured) are described by metadata to enable their discovery.

[3] The business reference model is one of the components of the Federal Enterprise Architecture (FEA) that is currently being used in the State.

- **Accessibility -** Making both the critical and non-critical data easily accessible to users and applications would foster enhancement to existing applications and to the development of new applications. This in turn would enable better services for the public and improve decision making for the Legislature, State agencies, and local governments for the services provided to the public. The users and applications will post or retrieve data to or from a "shared data space." The access to the data assets is made available to any user or application as permitted by policy, regulation, or security. For more information on policies, regulations, and security around sharing of data assets, refer to National Institute of Standards and Technology (NIST) Publication 800-100 Information Security Handbook, Chapter 6[v].
- **Standardization -** Standardized data approaches are incorporated into COI's/LoB's process. Once the standardized data approach is incorporated into the COIs/LoBs, this approach will then trickle down to department processes and practices. The benefits of enterprise and community data would be prevalent throughout the State.
- **Comprehensibility -** By making the data comprehensible, both structurally and semantically, users and application developers can readily determine how the data may be used for their specific needs.
- **Security -** By making the security level of each data asset available, users and applications should be able to determine and assess the authority of the source.
- **Interoperability -** In the cases where systems through interfaces have many-to-many exchanges of data, metadata would allow mediation or translation of data between interfaces, as needed.
- **Availability -** Data must be highly available with near zero downtime to enable mission critical systems.
- **Harmonization -**To ensure quality and accuracy of data, harmonization of common data must be accomplished. Example: Person, Location.
- **Responsive to Users Needs -** To ensure satisfaction, the perspectives of users, whether data consumers or data producers, are incorporated into data approaches through continual feedback.
- **Evolutionary -** Taking the complexity of the State's IT environment into account, the solution must be incrementally implementable as well as adaptable to meet the needs of the business as the State's business evolves.

3.3 Design

A strategy is a well thought out plan, and for the Statewide Data Strategy we visualize "a desired future IT state" for the State of California based on the CDS standards-based platform. In addition to the design principles described in Section 3.2, the CDS platform will embrace the following fundamental concepts:

- Data integrity
- Data ownership
- eDiscovery

- Readily assessable
- Standard interfaces
- Supporting business processes
- External partner support
- Secure information sharing
- Auditable transactions

This vision is predicated on three key elements:

1. Creation of a "Shared Network Cloud" for data tagging, sharing, integration, aggregation, searching, and retrieving.
2. Creation of a "California Trading Partner Network Services" to enable State agencies to interface with trading partners, such as federal agencies, towards information exchange.
3. Overall governance to ensure delivery of expected results based on well-defined business goals and to manage the design, deployment, security, and audit of services.

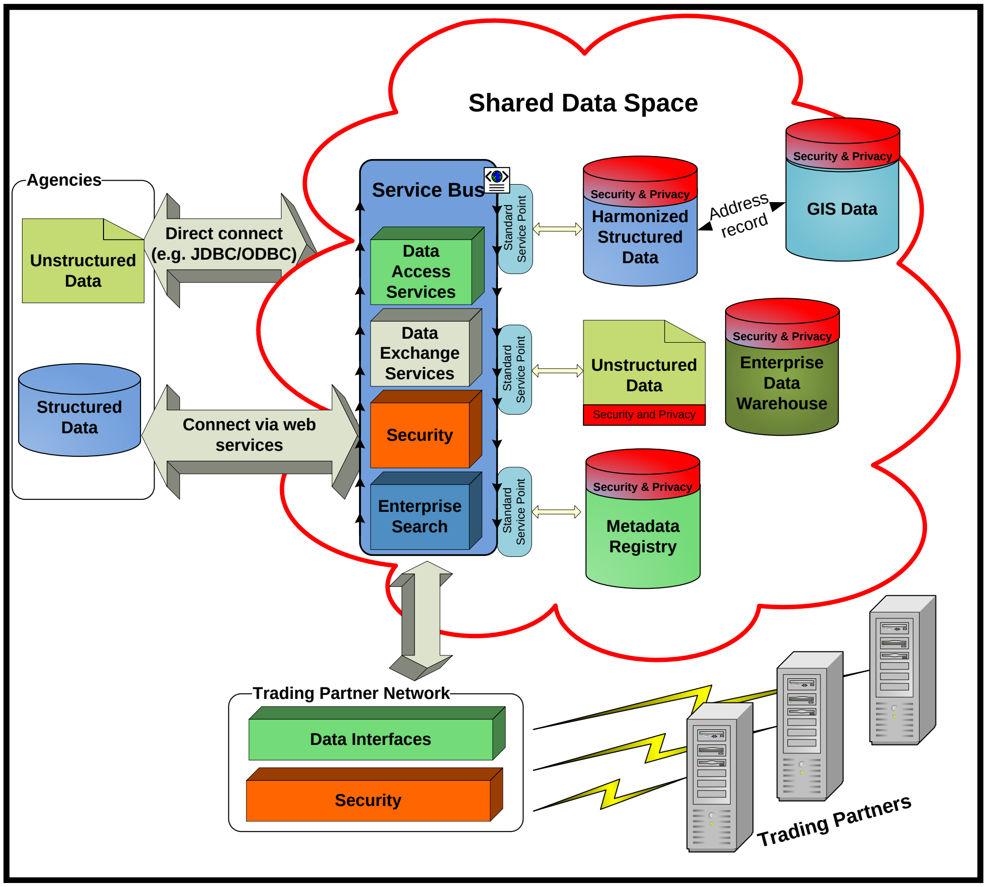


Figure 3-1 – California Data Services Conceptual Architecture

3.3.1 Shared Network Cloud

A “Shared Network Cloud” provides the foundation of the CDS platform. The CDS platform provides the foundation for the storage and sharing of data through a standard set of services. CDS will also provide virtual or physical access to any number of data assets (e.g., databases, document storage, and registries). Any authorized user, system, or application that posts data would have access to the network.

At the core of the environment is the *Shared Data Space*. Shareable data is defined to be data that is either used or generated by Communities of Interest (COI) / Lines of Business (LoB) and is accessible by authorized users within or across the State. Shareable data could cut across existing organizational boundaries. An example would be multiple departments within or across agencies contributing to a business sub-function. In this context, data implies all data assets such as system files, databases, documents, official electronic records, images, audio files, web sites, and data access services. The vision could be achieved by populating a set of shared data assets with all shareable data and allowing any user or application to draw from this shared resource pool. The shared resource pool changes the paradigm from “process, exploit, and disseminate” to “post before processing”. Shared data is advertised and available for users and applications when and where they need it. Users and applications would search for and “pull” data as needed. Alternatively, users receive alerts when data to which they have subscribed is updated or changed (i.e., publish subscribe). Authorized users and applications have immediate access to data posted to the network without processing, exploitation, and dissemination delays. Users and applications “tag” data assets with metadata, or data about data, to enable discovery of data. Users and applications post shareable data assets to “shared” space for statewide use.

In the following sub-sections we will discuss the strategy towards persistence and access of data and consolidated warehousing. The access has been further broken down into Metadata and Service based models.

3.3.1.1 Harmonized Structured Data

Harmonized structured data, also known as master data, is a key component to the strategic vision. It normalizes critical shared data and supports the business rules to maintain the data. The benefit of the master data repository is that there is a single source of truth. Whether it be for a person, a business, or a location, a record in the master data repository represents the single accurate set of attribute values. The data is validated, duplicates removed, business rules enforced, and transactions audited.

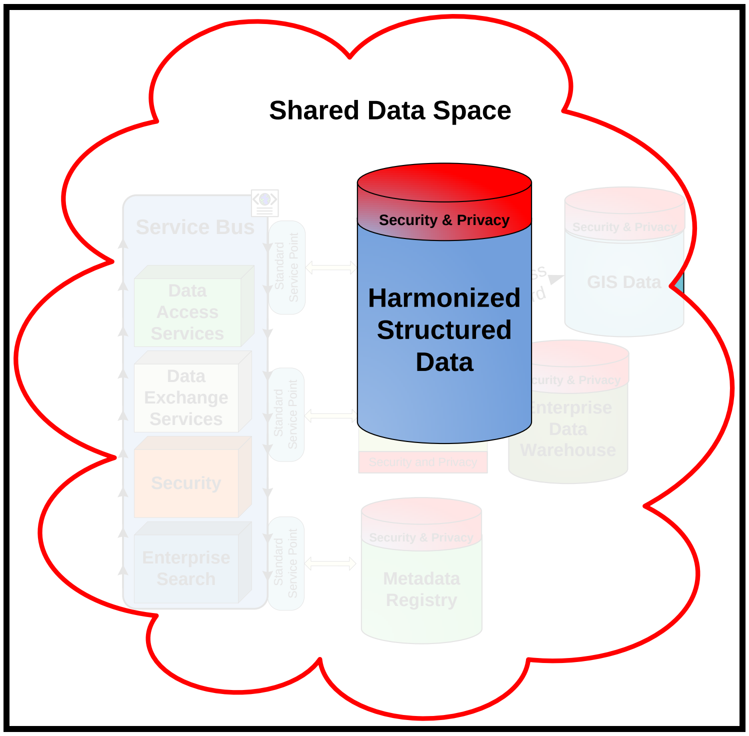


Figure 3-2 – Master Data

3.3.1.2 Unstructured Data Repository

Throughout the State of California, many document management, enterprise content management, and record management systems exist, each with its own set of business rules. A comprehensive, centralized enterprise content management system is recommended to support the sharing of unstructured data. Unstructured data is data that is not easily used programmatically. Examples include images, word documents, spreadsheets or sound files (e.g., WAV file). It should be mentioned that the term “unstructured data” is an imprecise term. Advances in technology improve the ability to analyze this unstructured data, blurring the lines between structured, semi-structured and unstructured data.

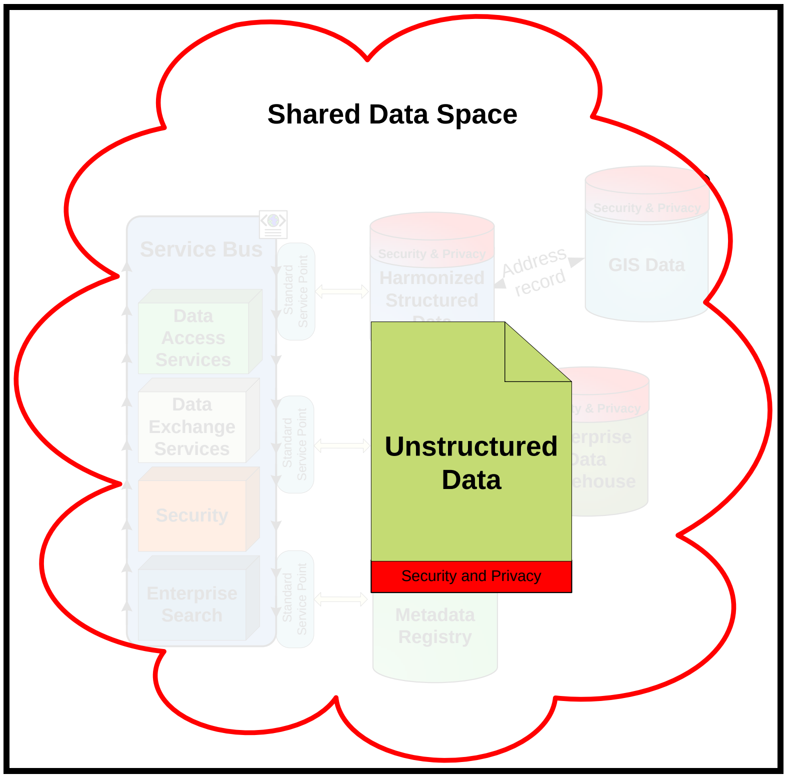


Figure 3-3 – Enterprise Content Management

3.3.1.3 Metadata Based Access

The cornerstone of this architecture is metadata. Metadata is data about data. It can be employed in a variety of ways to enhance the value and usability of data assets. Traditionally it is used to define data structures and relationships to support development of applications. In the case of California Metadata Registry, it would additionally enable discovery of data assets. Users and application developers will be able to quickly discover data assets by searching the registry.

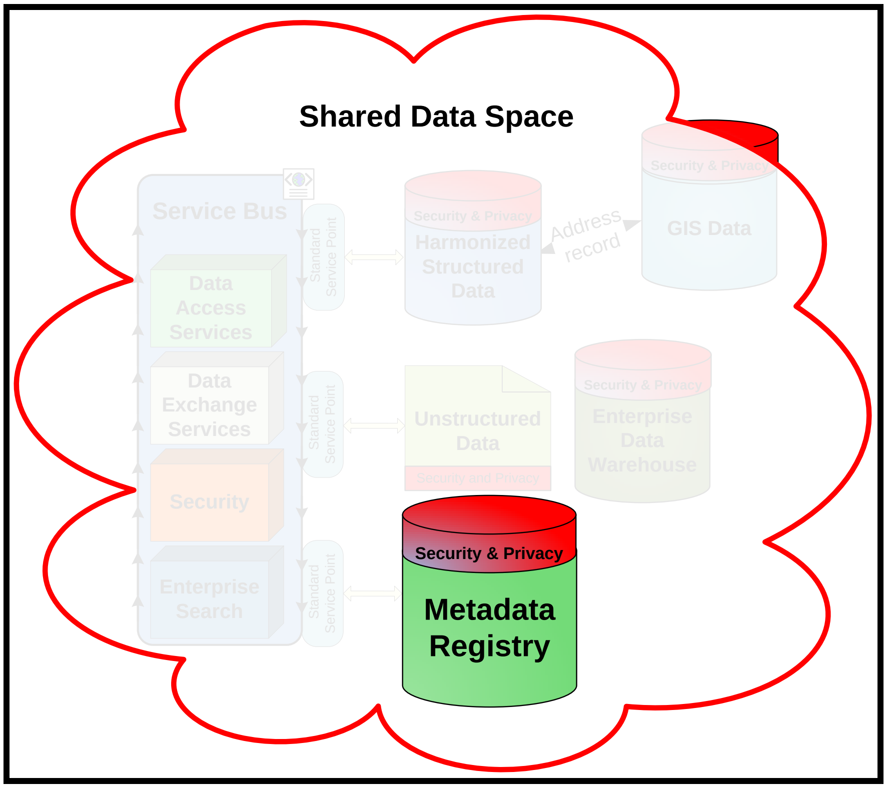


Figure 3-4 – Metadata Registry

There are many other types of metadata including vocabularies, taxonomic structures used for organizing data assets, interface specifications, and mapping tables. CDS capabilities use metadata in its various forms to support data asset discovery and interoperability and to provide a richer semantic understanding of all data and metadata.

The State should consider the use of an International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) 11179 standards-based metadata registry, California Metadata Registry, for storage and discovery. A metadata registry is a system that contains information that describes the structure, format, and definitions of data. Typically, a registry is a software application that uses a database to store and search data, document formats, definitions of data, and relationships among data. An ISO/IEC 11179 standards-based metadata registry additionally supports processes towards managing items. System developers and applications are the predominant users of a metadata registry. The Enterprise Architects for COIs/LoBs would be responsible for managing the registration process.

3.3.1.4 Service Orientation

In Figure 3-5, the primary unit of work is a ***service***. Access to data assets is enabled via services. Validated users, systems, or applications could search and discover access services that are available for the underlying data assets. The implementation for this model is predicated on Service Oriented Integration (SOI). SOI is an approach for architectures based on the concept of a service and brings the benefits of loose coupling and encapsulation at an enterprise level. Every COI would participate in governance and management of data services. While the representatives of a COI participate in decision making and data governance, the implementation and management would be handled by Enterprise Architects and Data Stewards for a COI.

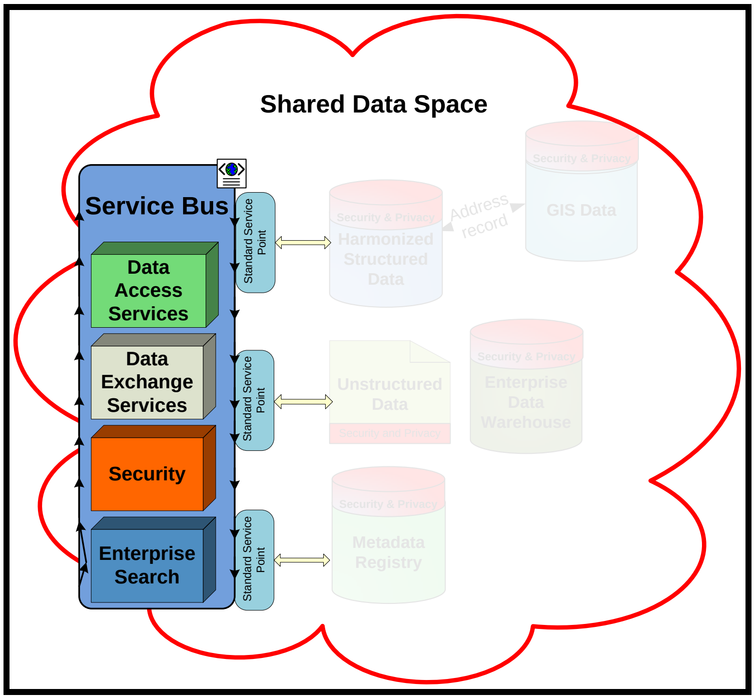


Figure 3-5 – Service Orientation

3.3.1.5 Consolidated Warehousing

As data is centralized and harmonized, it becomes much easier to use for analytics. The business rules for normalizing[4] the information across the agencies is required to harmonize the data in the Master Data Repository. So as more and more data is added to the Master Data Repository, the analytics that are normally performed at the agency level can be migrated to the shared space. The enterprise data warehouse in Figure 3-6 is simply a dimensional view of the normalized data from the Master Data Repository. There are no hard and fast rules that exist, but if the data is primarily sourced from the Master Data Repository, then much of the data consolidation and data cleanup work will have already been performed. Data from the agencies can also be added to the enterprise data warehouse or a data mart to support the analytics, but care must be taken to ensure the data is properly sourced, ownership rules applied, and update rules enforced.

[4] Normalizing data is a way of organizing the data structure so that it is suitable for general-purpose querying and free of insertion, update, and deletion anomalies. Data duplication and ambiguities are removed to improve the integrity of the data.

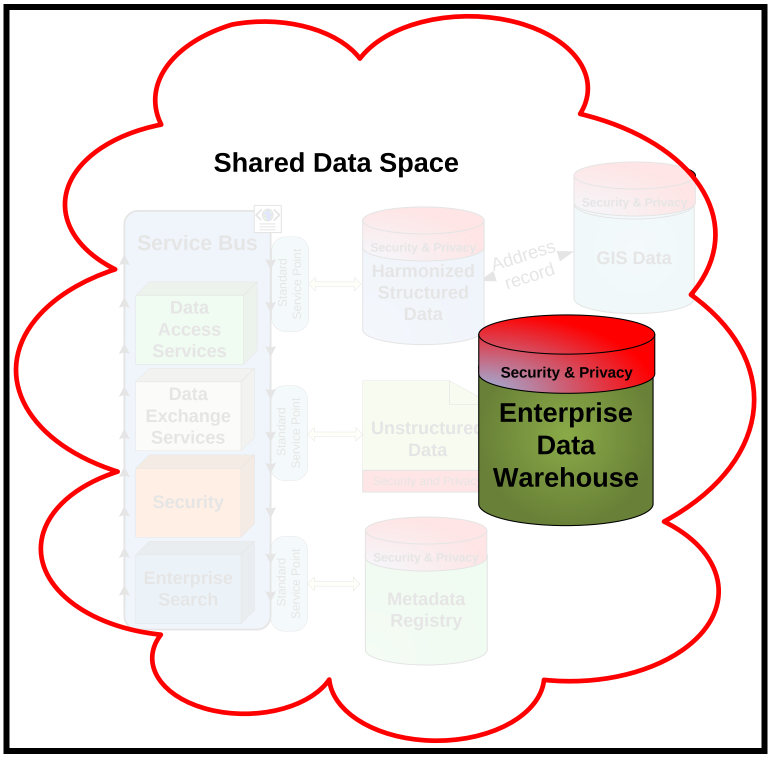


Figure 3-6 – Enterprise Data Warehouse

3.3.1.6 Geospatial Data

Geospatial data is part of the shared data solution. Structured Geospatial Information System (GIS) data should be contained in the Shared Space as part of its own repository. A common GIS repository, as part of the California Spatial Data Infrastructure, is an effort currently being led by the California Geospatial Information Officer (GIO). It is the goal of the GIO to have the framework data layers contained in the shared data space as data services for common use.

Geospatial data is data with a spatial component. Nearly all data is data about a place, but not all data is spatially enabled (i.e., is in a GIS format, and/or has fields with spatial definitions). The data strategy includes recommending an approach where GIS data is service based. Moreover this service returns geographic attributes through the structured data. For instance, the shared data space would house an address validation service and an address geocoding service. The address validation service would return a standard address and validate that the address is correct. The shared data space would then send the standardized address to a geocoding engine which would return the X and Y coordinate for the address. The X and Y location will be returned in the California standard projection and datum.

Additionally, the X and Y coordinate for the address could be sent to the full spatial data library to return any other requested framework layer attribute about that location (e.g., the US National Grid value, the Assessor Parcel Number, or the land use).

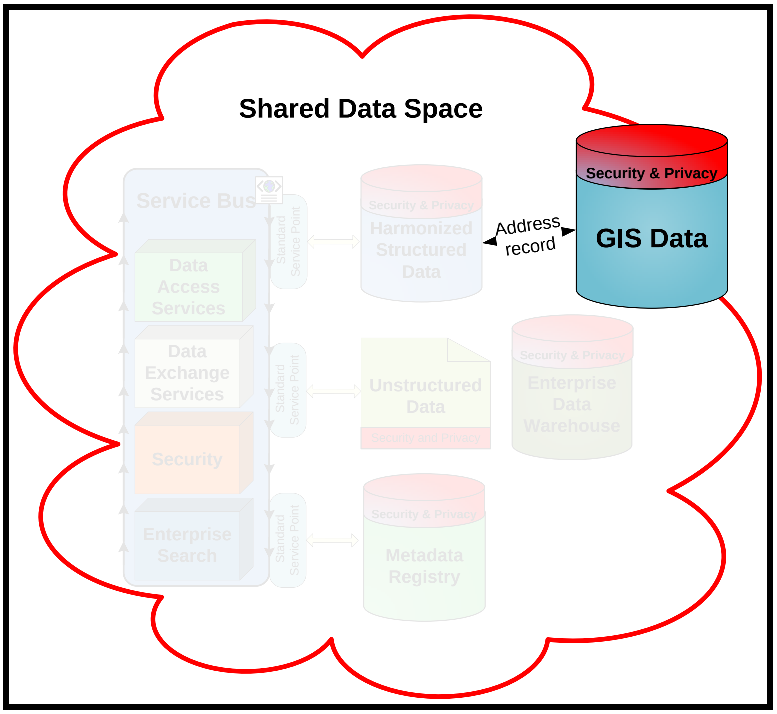


Figure 3-7 – Geospatial Support

3.3.2 California Trading Partner Network Services

The California Trading Partner Network Services (CTPNS) is a safe interface to expose to outside vendors and partners with semi-trusted relationships. A trading partner can be any entity that exchanges information with the State that is not a part of the State government systems. Examples include the federal government, local governments, and vendors. This gateway provides a means of sharing data without exposing the Enterprise Service Bus (ESB) to outside entities. It improves the security of the transactions and provides a means of managing the transactions from an outside source. An example of this is if a partner floods the trading partner network with transactions, it does not have to impact performance on the enterprise service bus. Business rules can exist that limit and control the transactions from external partners. These business rules will be described in the Trading Partner Agreement (TPA) for each trading partner.

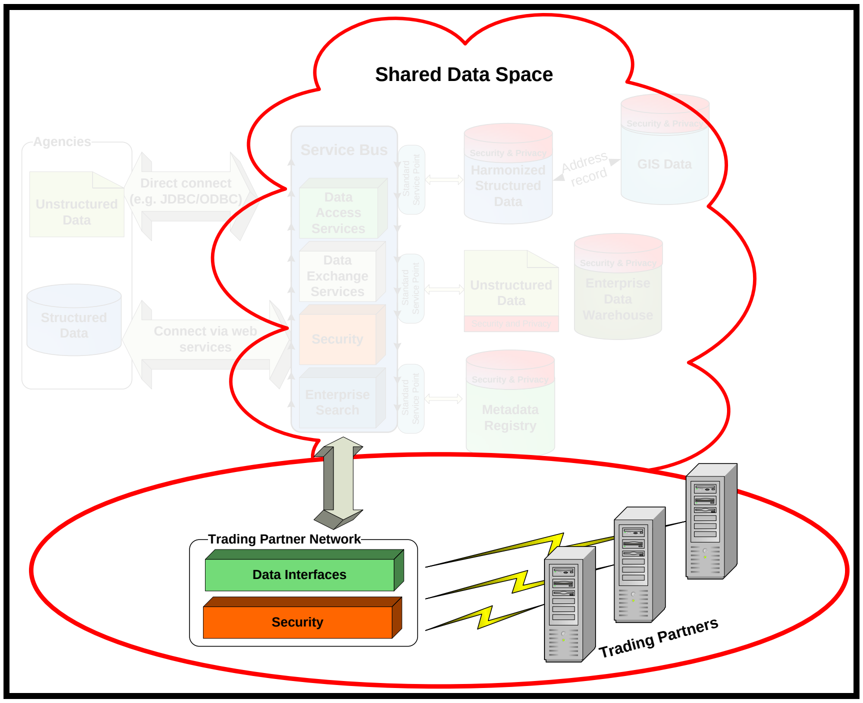


Figure 3-8 – Trading Partners

3.3.3 Governance

There will be a need for strong governance, first identifying and then managing shareable data. The management aspect encompasses the area of standardization and control of data elements. For a complete discussion on governance, refer to Section 5.

3.4 Architecture

To complement the goals and objectives recommended for CDS, we need to apply the Federal Enterprise Architecture (FEA) Guidelines in conjunction with the Data Reference Model (DRM). The FEA DRM is a flexible, standards-based framework. This standards-based framework will enable the State to reuse shared information across the State. The focus of this framework is on visibility and accessibility of data via the standard description and discovery of common data. In addition, it promotes uniform data management practices as well as standardization and control of data elements, definitions, and structures across COIs/LoBs. The framework of this focus is not isolated to the new architecture models, but also pertains to all legacy and new data assets, such as system files, databases, documents, electronic records, images, audio files, web sites, and data access services in the agencies. The use of this approach will improve flexibility in data exchange and support interoperability between systems without requiring predefined, pair-wise interfaces between them. This flexibility will be essential to the “many-to-many” exchanges in the environment.

While agencies will continue to manage and collect data using their existing systems, CDS will increase the potential for many other systems to leverage the same data without having to anticipate this use in the development cycle. For example, tightly engineered and real-time systems can offer “exposure” services that work “behind the scenes” collecting real-time data, storing it, and providing access and discovery through an enterprise interface. Exposure services can be designed to have little or no effect on performance-critical processes or predefined interfaces and still provide access to their data to unanticipated users. In an environment in which systems are continually being developed, deployed, migrated, and replaced, making allowances for unanticipated interfaces is essential.

The DRM provides value for agency data architecture initiatives by:

- Defining a means to consistently describe data architectures: The DRM's approach to Data Description, Data Context, and Data Sharing enables data architecture initiatives to uniformly describe their data artifacts, resulting in increased opportunities for cross-agency and cross-COI interactions.
- Bridging data architectures: The DRM provides guidelines to facilitate communications about data and data architecture between enterprise and data architects in their efforts to support the business/mission needs of the COIs that they support.
- Facilitating compliance with requirements for data architectures: The DRM's areas of standardization provide a foundation for agency data architecture initiatives and support the Federal Enterprise Architecture (FEA) Guideline's other reference models[5]. The Data Reference Model (DRM) puts forth requirements that can result in increased compatibility between agency data architectures.

3.4.1 Data Reference Model

The FEA Guidelines specify five reference models[5]. The DRM is one of those reference models, and comprises three major components: Data Description, Data Context, and Data Sharing (some of the information has been extracted from the DRM in Appendix D). These components and how they fit into the California Data Strategy are described below.

3.4.1.1 Data Description

Data Description provides a means to uniformly describe data by revealing patterns, relationships, subgroups, and exceptions in the data. The more ways you look at your data, the more fully you will understand the implications of the data. Enterprise Architects within the COIs are responsible for Data Description. In order to identify entities and designate metadata describing the various Business Sub-Functions within a COI, the Enterprise Architects will work with subject matter experts (SME's) from these Business Sub-Functions to decompose the data using the DRM Extensible Markup Language (XML) Schema. The decomposing process of identifying entities is part of the analysis to identify which data supports what aspects of a LoB, and how that data would help, where applicable, in the harmonization of the data. When the Data Description artifacts are developed with high quality standards, they support a COI's data architecture and enable Data Sharing services.

[5] FEA specifies the following reference models; Performance Reference Model (PRM), Business Reference Model (BRM), Service Component Reference Model (ScRM), Data Reference Model (DRM) and Technical Reference Model (TRM).

The California Metadata Registry, discussed in Section 3.3.1.3, would represent a "one-stop shop for developer data needs", and is a key component in achieving the COI's interoperability goals. All document formats, interface definitions, and exchange models used by systems will be stored in the California Metadata Registry. Developers can discover these metadata assets and utilize them to read, write, or exchange data that is made available statewide. All COIs have a responsibility to support interoperability through active participation in the California Metadata Registry. The California Metadata Registry will provide capabilities to further support interoperability through the use of translation and mediation services and for the sharing and reuse of processes. For example, a COI may develop and share a process for validating a mailing address. This process will be available statewide, and its associated metadata (input/output format and connection information) will be registered in the California Metadata Registry. Through this capability, the California Metadata Registry is more than just a simple repository of data formats**,** it is a comprehensive source for supporting design, development, and execution of processes (e.g., business logic) in a centralized, services-based data environment. The California Metadata Registry will facilitate the following Data Description capabilities.

- **Data Discovery -** The capability to quickly and accurately identify and find data that supports mission requirements.
- **Data Reuse** - The capability to increase utilization of data in new and synergistic ways in order to innovatively and creatively support missions.
- **Data Sharing -** The identification of data for sharing and exchange within and between agencies, departments, federal, state, and local governments, and other COIs, as appropriate.
- **Data Entity Harmonization -** An enhanced capability to compare data artifacts across government through a common, well-defined model that supports the harmonization of those artifacts and the creation of "common entities".
- **Semantic Interoperability -** Even when using service-oriented architectures or business process modeling approaches, one must contend with problems with different contexts and their associated meanings. Semantic interoperability is a capability that enables enhanced automated discovery and usage of data due to the enhanced meaning (semantics) that is provided with the data itself.

3.4.1.2 Data Context

Data Context is any information that provides additional meaning to data to correlate the data to the purposes for which the data was created and will be used. The use of a data context facilitates discovery of data through an approach of categorizing data according to taxonomies. Additionally, data context enables the definition of authoritative data assets within a COI. It is the information that makes it possible to provision a Context Awareness Service to support a COI or collaboration among COIs. Within a Context Awareness Service, one identifies the existence of a Data Asset and enables a user to discover whether it is potentially relevant to a given information need. The service makes Data Context artifacts, developed in accordance with the Data Context section of the DRM abstract model, available for use. These artifacts are chosen by the COI to reflect government related business needs and contain adequate information to support government related decision making.

We recommend defining a single common taxonomy, California BRM Taxonomy, for all COIs. As the name suggests, it corresponds to the context of California BRM. It is a hierarchical taxonomy with Business Area, Line of Business, and Sub-function topic as the three levels of the hierarchy. An Enterprise Architect for a COI is responsible for compiling taxonomy for the COI by speaking with both business process experts and database administrators. Additional taxonomies are recommended and defined based on organizational structure, subject areas, data asset types, etc.

3.4.1.3 Data Sharing

Data sharing is the use of information by one or more consumers that is produced by a source other than the consumer. It supports the access and exchange of data, where access consists of ad-hoc requests (such as a query of a data asset), and exchange consists of fixed, re-occurring transactions between parties. Data sharing needs are often difficult to predict in advance. At a broad level, data can be shared in two ways, through Data Exchange Services and through Data Access Services.

Data Exchange Services are services that allow movement of large amounts of data between repositories whereas *Data Access Services* can provide the detail related to the data itself or provide a single purpose service to provide an aggregate function. For example, data trends, common metrics, or common summary information could be provided as a data access service enabling the agencies to have access to this information on demand. The complexity of the aggregation algorithm would be hidden within the service itself.

3.4.1.3.1 Data Exchange Services

Data Exchange Services are primarily used for transfer of data between repositories. The various types of data exchange services identified in the DRM are:

- ***Extract, Transform, Load:*** In these services, structured data from a data source (the extract) is first converted to match structured data required by a target database (transform). Finally, the target database is updated with the transformed data objects (load).
- ***Publication:*** In these services, a document is assembled into a desired format by aggregating structured data or documents.
- ***Entity/Relationship Extraction:*** In these services, facts from unstructured documents are pulled out to form structured documents or data objects.
- ***Document Translation:*** In these services, a document from its original format is transformed to a required format.

3.4.1.3.2 Data Access Services

Data Access Services provide other services access to data. The various types of data access services identified in the DRM are:

- ***Context Awareness Services:*** These services allow the users of a collection to rapidly identify the context of the data assets managed by the COI.

- ***Structural Awareness Services:*** These services allow data architects and database administrators to rapidly identify the structure of data within a data asset (i.e., a structural awareness services make the Data Description as defined within the DRM available for use).
- ***Transactional Services:*** These services enable a transactional create, update, or delete operation to an underlying data store while maintaining business and referential integrity rules. These services allow external services or end users to execute data related functions as a part of a workflow or business process.
- ***Data Query Services:*** These services enable a user, service, or application to directly query a repository within a collection.
- ***Content Search and Discovery Services:*** These services enable free text search or search of metadata contained within the documents in a repository.
- ***Retrieval Services:*** These services enable an application to request return of a specific document from a repository based upon a unique identifier, such as a URL.
- ***Subscription Services:*** These services enable other services or end users to automatically receive new documents added to a repository in accordance with a predetermined policy or profile.
- ***Notification Services:*** These services enable other services or end users to automatically receive alerts of changes to the content of a repository in accordance with a predetermined policy or profile.
- ***Auditing Services:*** These services enable auditing of sensitive transactions to support the monitoring of data usage.

3.4.2 Shared Data

The shared data illustrated in the Shared Network Cloud in Figure 3-1 has several architectural components that must be discussed in detail. Shared data can be of two types:

- Structured
- Unstructured/Semi-Structured

The structured data can be of two varieties as well, master data and dimensional reporting data. Once a 'master' copy of the data is instantiated, this impacts the integration approach, the warehousing approach, and the approach to governing the data. The dimensional data is the Master Data restructured to better support reporting.

One of the more powerful features of this shared data model is the ability to relate information across multiple domains (e.g., agencies or departments). The challenge in sharing data lies in organizing the data in a consistent manner so that it is meaningful in each domain. Once that is done, patterns of usage across agencies and departments emerge. This challenge is illustrated in the following real-life example: the difficulty of correlating data related to services offered at a hospital to the data on services received by patients. At initial glance, one might believe this would be an easy match, but this is not so. Services offered by the hospital may be driven by insurance reimbursement structures with details on usage of different hospital resources, e.g. operating rooms, whereas patient information may include the listing of procedures performed and medications administered. Since two different systems store this data very differently, it is difficult to build a relationship between the two.

Another cross-agency example involving addressing is identifying all addresses that a person provides to State agencies at any given time. For example, people receiving unemployment benefits may give a different address to the Employment Development Department (EDD) than to the Franchise Tax Board (FTB), since EDD provides unemployment benefits and FTB collects taxes. It is important to tie both addresses to the same person for proper accounting.

When more data is shared and the relationships within the data are understood, the State will have a view of a person or organization that cuts across agency IT systems. This accurate and comprehensive view provides the means of seeing the 'big picture'. It allows the State to consider data trends that exist between data that is maintained by different agencies, allowing the State to answer those formerly unanswerable questions about the data.

3.4.2.1 Master Data

The implications of having a single consistent source of data are fairly far reaching. The term 'single consistent' implies no duplication of data (i.e., only one copy of data is maintained). This leads us to a normalized data model that has the data represented only once. This architectural approach can be described as a 'Master Data' data store.

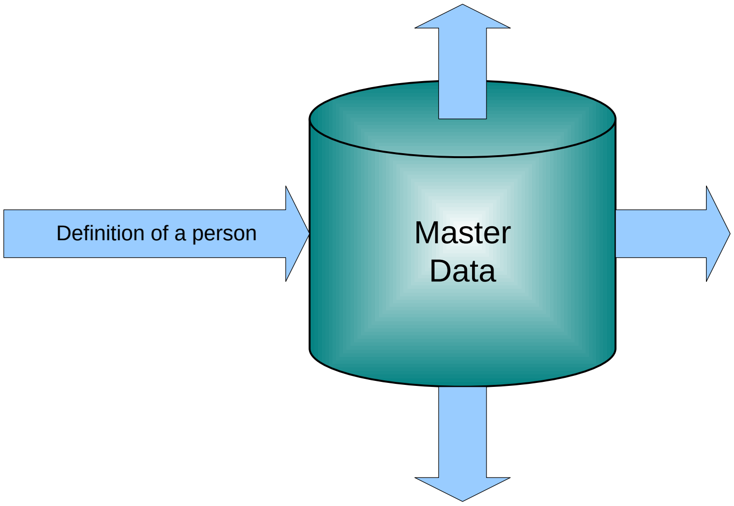


Figure 3-9 – Master Data Concept

The concept is simple. Data is extracted and organized so that only a single copy of data exists. The rules around data ownership, data maintenance, and data access are established in both the database design and in the application design. Data in this structure caters to Online Transaction Processing (OLTP) but does not perform well for all types of reporting, and therefore a transformation to a dimensional model would be necessary to support data warehousing needs.

The reason why this approach works so well in industry is that it addresses one of the current issues at hand, lack of one version of the truth. The State has the same challenge: a single, comprehensive version of the truth is needed. With very few exceptions, each agency currently maintains “its own” data. Very little is shared across the agencies. In addition, none of the agencies have a complete view of the data. For example, one agency may have the person’s income, taxes paid, and address, whereas another agency may maintain a list of services provided to the person. In this example, to get a comprehensive view of the person, information across agencies is required.

3.4.2.1.1 Data Integration Overview

Every agency uses its own set of applications in so called silos. Most of them have minimal integration to the other agencies. Figure 3-10 illustrates the agency IT systems in their most basic state[6].

This diagram is only for illustrative purposes. Each of the agencies that are listed are quite complex in and of themselves. Each agency contains multiple departments (or divisions), and each of these departments use many applications. Most of these applications are on disparate platforms and use a widely different set of technologies. These departments in reality contain their own set of complex data silos.

[6] The diagram has been simplified for illustrative purposes. Although the figure implies no data sharing across the agencies, there is currently some data sharing that exists across the State agencies but it is not comprehensive.

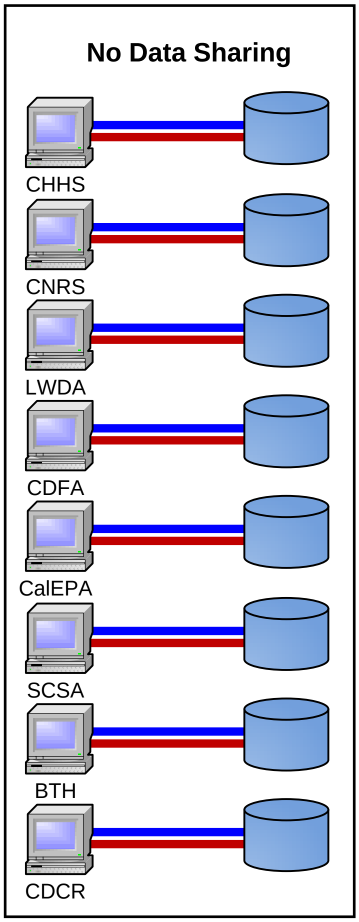


Figure 3-10 – No Data Sharing

As interfaces are needed, a point to point interface is created between applications. Most of these interfaces are specific to the need and functionality is duplicated[7].

As these point to point interfaces are created, they have to be maintained. These interfaces may have to be modified as application enhancements are made. The sheer volume alone can be overwhelming. In addition, they tend to be tightly coupled with the interfacing applications, and therefore have a tendency to be fragile. Availability of an application is only as good as its weakest link.

From the following figure 3-11, one can see that the point to point interface approach can quickly become overwhelming.

[7] For example, one California department has 47 interfaces to the IRS, many of which are very similar.

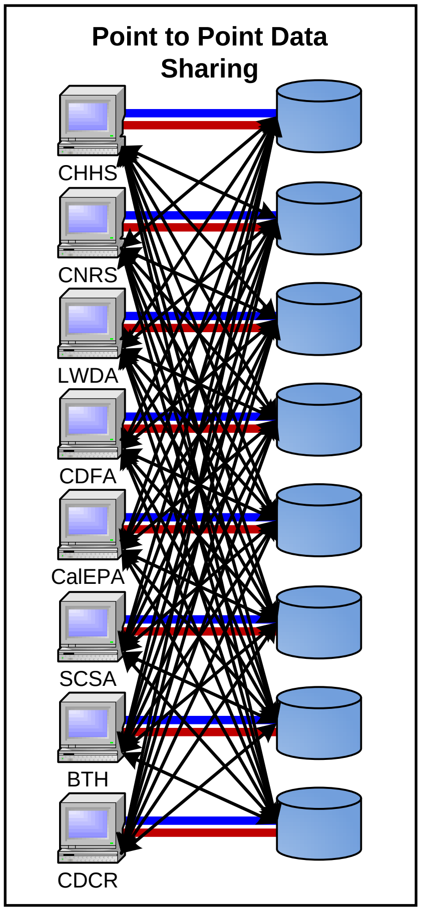


Figure 3-11 – Data Sharing by Point to Point Interfaces

Traditional Enterprise Application Integration (EAI), illustrated in Figure 3-12, hides the complexity within a service bus infrastructure, which is expensive to implement for a single agency and not viewed as feasible across multiple agencies.

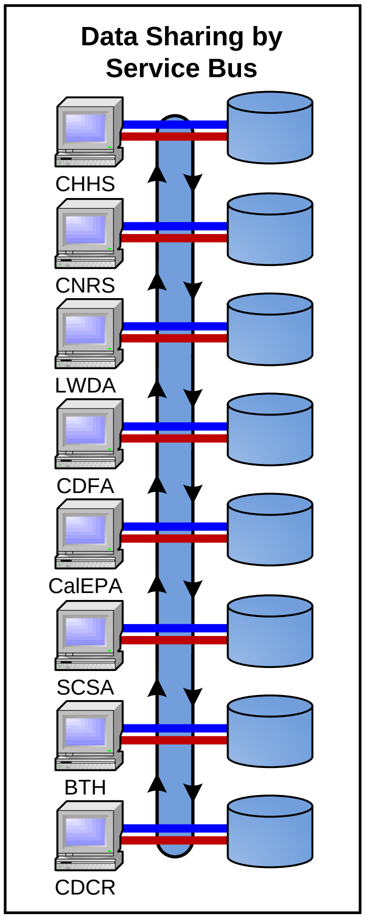


Figure 3-12 – Data Sharing by Service Bus

Traditional Enterprise Data Warehousing, illustrated in Figure 3-13, collects data from the silos, using Extraction, Transformation and Load (ETL), which is easy to implement, but provides a mostly read only view; its usefulness is limited to reporting and analytics.

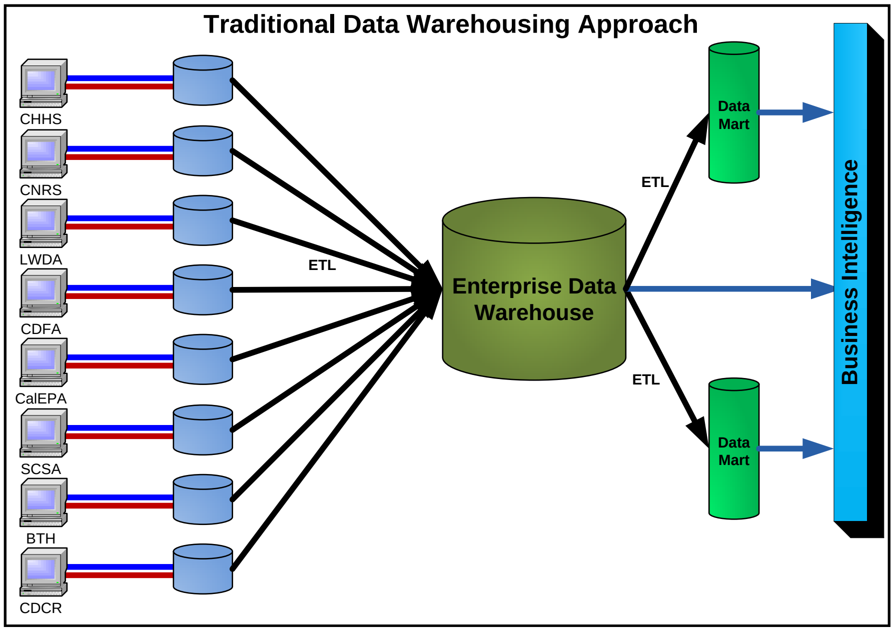


Figure 3-13 – Enterprise Data Warehouse Traditional Approach

Master Data Management (MDM), illustrated in Figure 3-14, leverages the successful ETL approach to generate a centralized master data repository, and employs a "reversed ETL" approach to disseminate the cleaned up data back to the silos, while maintaining the ability of reporting and analysis.

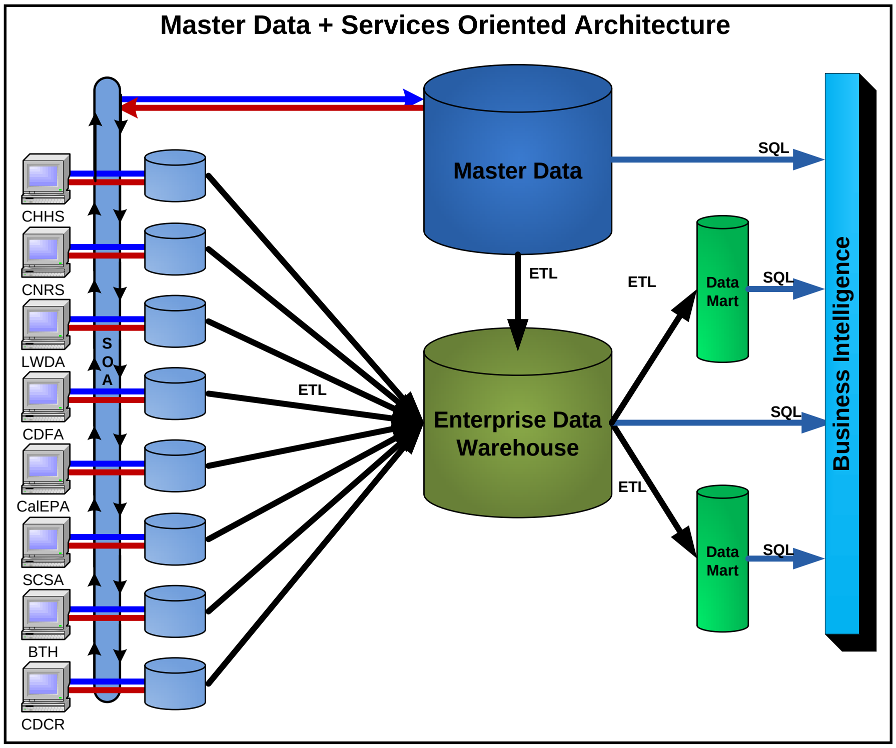


Figure 3-14 – Enterprise Data Warehouse Master Data Approach

3.4.2.1.2 Master Data Management

Master Data represents the business data that is shared across more than one transactional application. Since this data is involved in transactions, it must be structured like a normalized transactional database, an OLTP database structure. In addition to maintaining this structure, dimensions are added to the data to support analytics. This structure maintains a single source of truth across the IT enterprise. It is also important to note that since MDM supports transactional applications, it must support high volume transaction rates.

3.4.2.1.3 Master Data Management Characteristics

- Has a flexible and easily extendable data structure that contains a single copy of each data object that it supports. The data structure supports OLTP transactions and is independent of any application.
- Maintains information about the data in a metadata repository.
- Supports security and ability to search both the data and the metadata.
- Supports complex data ownership and update rules.
- Supports a data quality interface to assist in the cleanup of duplicate data and the maintenance of data.
- Has a triggering mechanism to notify and post changed data to connected systems.
- Has a comprehensive auditing and history capability to track all changes.
- Uses a single platform to manage all master data in order to prevent contributing to new silos of information.
- Provides an analytical foundation for analyzing master data.
- Uses a highly available and scalable platform that supports heavy mixed workloads.

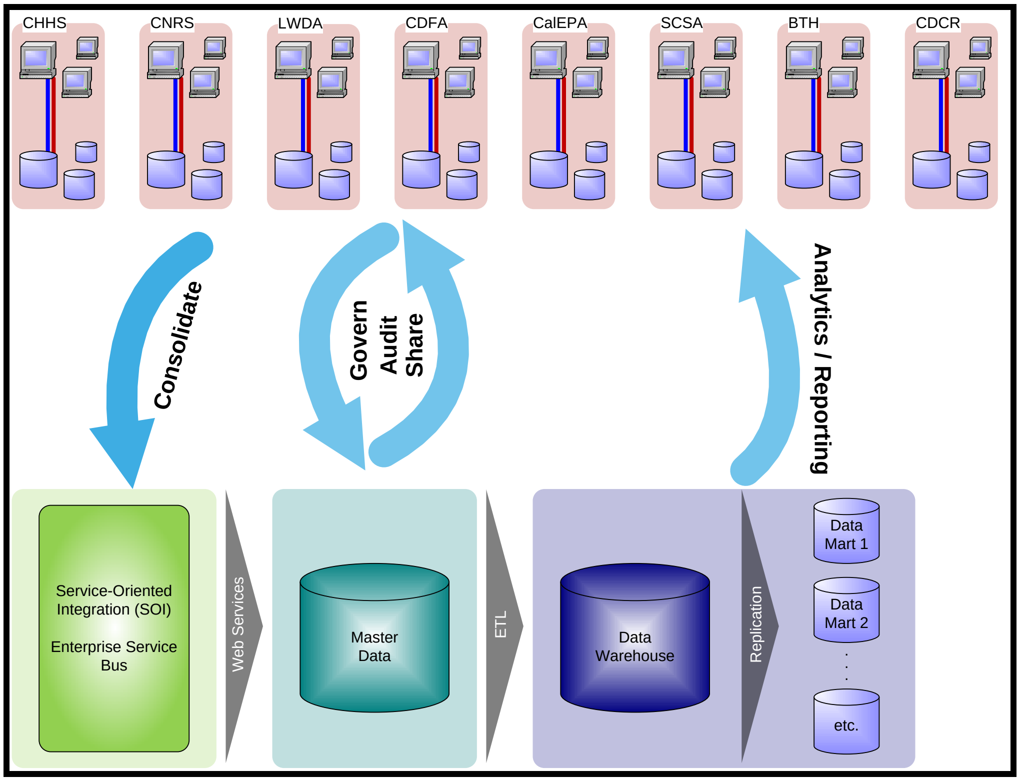


Figure 3-15 – California Data Services - Agency Interaction

Agencies interact with this solution as follows:

Consolidate - Updated data is pushed from the source systems to the Enterprise Service Bus (ESB) and is applied to the Master Data. Update logic and workflow ensure that the data is consolidated properly. Web services interface the existing systems to the ESB and from the ESB to the Master Data repository.

Govern, Audit, and Share - Once data is placed in the Master Data repository, it can then be managed, audited, and shared with the other agencies.

- **Govern** - A part of governance is the data stewardship. The update rules must be maintained to determine who can update the data, who wins in case of an update collision (multiple messages may update the same logical record), and which details are updated and how.
- **Audit** - An important feature of the Master Data Management is auditing the changes made to the master data, who made these changes, and who even accessed this information.
- **Share** - Once data is in the MDM repository, it is available to share. The MDM repository is structured like a transactional database (i.e., the data is normalized). Attention is given to the design to make it very efficient with frequent yet small transactions. If a service so chooses, it can request the data from the MDM repository and be assured that that represents a recent master copy of the data.

Analytics and Reporting - The data can be transformed and de-normalized to accommodate a dimensional model. The main difference between a data warehouse and a data mart is that a data warehouse is designed to accumulate data, whereas a data mart is a smaller version with limited scope and depth for the data, and not meant for large scale accumulations. The beauty of data marts is that they can be destroyed and rebuilt as needed to support the business. For example, if a report needs to be expanded, and a new field needs to be added, the data mart can be simply dropped, rebuilt, and reloaded to accommodate that new field. In a data warehouse, since data is being accumulated, and the data volume is much larger, this is a bit trickier. The new field must be added, and some process to update the database must be created and validated.

3.4.2.1.4 Benefits

The upside to this architecture is significant. Since data is maintained centrally, objects (e.g., person, facility, or business) are well defined. Agencies interact with data only through well defined interfaces. The data is cataloged, and a data context is associated with the data that identifies its owner, creation date, and significance[8].

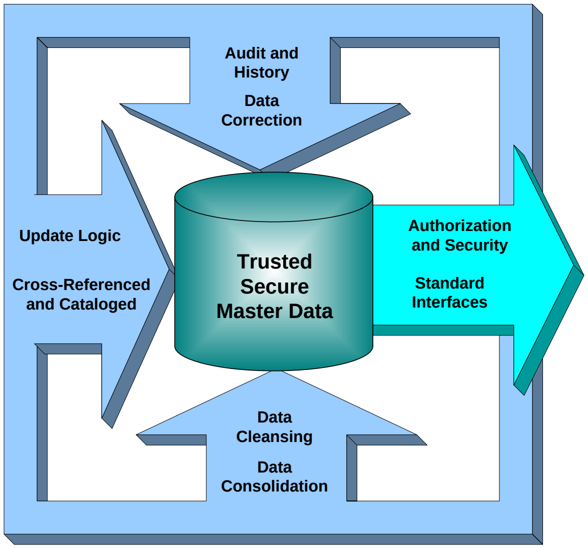


Figure 3-16 – Master Data Benefits

Standard Interfaces

It is important to note that one size need not fit all, as different technologies can be used for different interface types. Unlike point to point integration, only a few standard interfaces are created, and all systems must conform to the 'standard' to take advantage of the data. To support differing data needs, information can be transformed on the agency side of the interface.

Enterprise Search and the Metadata Catalog

Another benefit of centrally located Master Data shared data space is the ability to search across all agency data. Since data from these different sources are now readily

[8] Data Context will be discussed in detail in the data context section of the document.

available, complex trends between data can be better analyzed. An example could be the analysis of the relationship between income, taxes paid, and services accepted.

Aiding the search is the data context. Data can be in effect labeled as it is created within the data store. Data creator, data owner, transaction that created the record, creation date, data rights, and a title allow the cataloged information to be searchable. It also signifies the value or importance of the data to the State of California.

Well Defined, Validated Data Source

An intrinsic benefit to this approach is the implied guarantee that the Master Data is being carefully maintained to be the 'best source' for data. Ultimately, the data did not originate in the Shared Data Space, but there is an implied level of service in the design. Before data can be a part of the Master Data, it must be received via a well defined and secure interface, have duplicates removed, and be validated and cataloged.

Standards

The Master Data repository infrastructure must support high performing, highly available, and highly scalable configurations using industry standards. At a minimum, it must support the ANSI 92 SQL standard. It should also have native support for Java and JDBC, PHP, Perl, COBOL, ODBC, and other common industry standards and languages.

3.4.2.1.5 Challenges

With any alternative, there is work that needs to be performed to ensure that issues are avoided and risks are managed. The areas to consider are: Data Cleanup, Data Availability, and Security.

Data Cleanup

Probably the biggest challenge with this approach is reconciling the data. Formatting and validating addresses can be difficult since much of it is entered in a free format fashion in most applications. The same address can be entered differently within two applications making matching difficult.

Another issue to emphasize the point, the research revealed that there is no common unique identifier for a person. The most obvious choice would be Social Security number. Unfortunately, a Social Security number cannot be used by law, so there are few alternatives available. The risk of using Social Security number as a primary key was identified in "The HEW Report of 1973", and it was mitigated by addressing it in "The Privacy Act of 1974"[vi]. In some cases, people are defined within a department's systems multiple times and in different ways. A consolidation and data cleanup effort must be performed to allow definition of unique "person".

Other types of data like facility, location, and business have unique identifiers. They may be an easier initial target subject area to instantiate in the in the Master Data.

Data Availability - Once this central data repository exists and is being used, its availability will be critical. Any downtime would impact many applications, so its underlying infrastructure and interfaces must be very carefully designed and implemented. Both topics will be discussed in detail later in the strategy.

Security - Centralized and potentially highly sensitive data brings with it a higher security requirement. In addition to just controlling access, the transmission and storage of this data must also be encrypted. Security for accessing, transmitting, and storing this information is a requirement and is an integral part of the security component of this strategy.

Master Data Management (MDM) is a key component in the overall architecture. One benefit with this architecture is that the MDM can be grown in an evolutionary manner. As data is added to the MDM, so can web services be added or extended to support the additional data requirements. This alleviates the 'big bang' project lifecycle, and thus lowers the overall risk.

3.4.2.2 Unstructured Data

Agencies implicitly or explicitly generate and use a substantial volume of "Unstructured Data Assets" in addition to "Structured Data Assets", Documents of type MS Word, Multimedia, HTML, and PDF are good examples of such data assets. Every department has the capability of generating such information. We have observed one common pattern towards generating such data across many departments. It is the usage of attachments for Structured Data Assets. Without an effective means of capturing and organizing documents, the management of such data would become extremely hard. Without a centralized knowledge base, outdated documents cause confusion and can become difficult to eradicate. Consolidating all of the data assets in a secure repository can significantly reduce the amount of time spent on managing documents.

In order to establish a comprehensive data sharing strategy, you need to identify and make such data assets shareable too. To execute, we recommend the usage of a comprehensive solution that supports the following features:

- Records Management
- Metadata
- Indexing
- Retrieval

3.4.2.2.1 Records Management

Records management minimizes the risk associated with compliance, legal actions, discovery, and regulation by allowing you to accurately capture, identify, store, and dispose of business records properly. It further enables enforcing of document retention policies to meet the specific needs of business processes and ensure compliance with regulatory requirements.

Benefits

- Correctly identifies documents as records.
- Maintains compliance with regulations such as eDiscovery, HIPAA, PCI, and CPR Security.
- Manages lifecycle of records with an audit trail.

- Protects accidental or unauthorized alteration, destruction, or retention of records.
- Assigns records policies to folders and documents when they are created or imported.

3.4.2.2.2 Metadata

In the case of Structured Data Assets, the data is in a well defined structured format. However, it’s not the same when it comes to Unstructured Data Assets. They may or may not have context for data. Examples of data context include ***Title*** and ***Author***. Enterprise Architects may choose to specify context metadata fields that are appropriate for a business process. The system should enable storage of metadata either manually or automatically, so that context for data, where supported, can be extracted and stored in the system via standard or custom tools or manually, depending on frequency of use.

Benefits

- Documentation of data characteristics to enable sharing, discovery, retrieval, and exchange of data.
- Sets up common data standards between organizations. Exchange of data among organizations is facilitated with the common data standards.

3.4.2.2.3 Indexing[vii]

The system should enable automatic indexing of documents as they are added. To expedite the execution of search queries across a large set of documents, the documents need to be indexed. In the indexing stage, the system will scan the text of all the documents and build a list of search terms, often called an index, but more correctly named a concordance. In the search stage, when performing a specific query, only the index is referenced rather than the text of the original documents. The indexer will make an entry in the index for each term or word found in a document and possibly its relative position within the document. Usually the indexer will ignore stop words, such as the English "the", which are both too common and carry too little meaning to be useful for searching. Some indexers also employ language-specific stemming on the words being indexed, so for example any of the words "drives", "drove", or "driven" will be recorded in the index under a single concept word "drive".

Benefits

- Improved performance retrieving documents.
- Improved usability related to document retrieval, allowing for imprecise searches to retrieve documents based on words that sound like or are a derivation of the search criteria.

3.4.2.2.4 Retrieval[vii]

To complement Indexing, the system should support searches toward the retrieval of documents. It should support the following search features:

- **Keyword -** During indexing, a list of keywords is provided for each of the documents. Keywords describe the subject of the document, and could include synonyms of words that describe the subject. Keywords improve recall, particularly if the keyword list includes a search word that is not in the document text. Keyword search enables searching of a term from the supplied list of keywords.
- **Boolean -** A Boolean search is where a user specifies a relationship between any two or more search terms. Either both must be true (the AND condition), at least one must be true (the OR condition), the first must be true and the second false (the ANDNOT condition), that at least one must be true, but not both (the XOR condition) or that the two terms must occur near (within 5 words of) each other (the NEAR condition).
- **Phrase -** A phrase search matches only those documents that contain an exact match of a given phrase.
- **Proximity -** Proximity is a form of free text search where the proximity of two or more search terms is specified.
- **Field -** Field enables searching of terms that are present in the metadata associated to the documents.

Benefits

- Allows multiple means of finding a document.
- Improves usability as it is consistent with the functionality provided by popular internet search engines that are commonly used.

3.4.3 Metadata Based Access

To facilitate implementation of the architecture recommended in Section 3.4.2.1.3, we encourage creating XML documents representing Structured, Unstructured, and Semi-structured data assets and registering them with California Metadata Registry. The registry would enable users, systems, or applications to search and discover data assets. A sample XML document has been included in Appendix E.

World Wide Web Consortium (W3C) has defined a XML Schema, DRM XML Schema, which serves as an abstract meta-model for DRM. It represents the three major standardization areas (Data Description, Data Context, and Data Sharing) of DRM, thus facilitating its implementation. The latest version of the schema, Draft Version 4 that is available at W3C, is included in Appendix E. The schema will:

- Support the DRM's primary use case of facilitating a statewide information sharing
- Support statewide harmonizing of data artifacts, and establishment of authoritative data sources
- Provide an open and well-documented standard to enable organizing and categorizing of information in ways that are searchable and interoperable across the State

For a more detailed explanation regarding the concepts and standards pertaining to respective attributes specified in DRM XML Schema, please see FEA DRM in Appendix D.

3.4.3.1 Description

The following concepts comprise the Data Description standardization area and are taken from the Data Description Section of the DRM, provided in Appendix D[viii]:

- **Entity:** An abstraction for a person, place, object, event, or concept described (or characterized) by common Attributes. For example, “Person” and “Agency” are Entities. An *instance* of an Entity represents one particular occurrence of the Entity, such as a specific person or a specific agency.
- **Data Type:** A constraint on the type of physical representation that an instance of an Attribute may hold (e.g., "string" or "integer").
- **Attribute:** A characteristic of an Entity whose value may be used to help distinguish one instance of an Entity from other instances of the same Entity. For example, an Attribute of an “Organization” Entity (e.g., Business) may be “Tax ID”. A Tax ID, also known as an Employer Identification Number (EIN), is used to distinguish one business (i.e., one instance of an “Organization” Entity) from another.
- **Relationship:** Describes the relationship between two Entities. (e.g., person and house address)
- **Digital Data Resource:** A digital container of information, typically known as a file. It will be a container for the metadata about the data resource. A Digital Data Resource may be one of three specific types of data resources, each corresponding to one of the three types of data described below.
- **Structured Data Resource:** A Digital Data Resource containing structured data. Structured Data is data described via the E-R (Entity-Relationship) or class model. This data can be accessed in a uniform manner, independent of data values, once the Data Schema is known.
- **Unstructured Data Resource:** A Digital Data Resource containing unstructured data. Unstructured data is data that is not described according to an E-R model, but is a more free-form format, such as multimedia files or unstructured text.
- **Semi-structured Data Resource:** A Digital Data Resource containing semi-structured data. Semi-structured Data is data that has characteristics of both structured and unstructured data.
- **Document:** A file containing Unstructured and/or Semi-Structured Data Resources.

The State should consider ***metadata descriptions*** for Structured, Unstructured, and Semi-structured data and register them with California Metadata Registry. Users, systems, or applications could search and discover the data assets via the registry.

For a more detailed explanation regarding the concepts and standards pertaining to respective attributes specified in DRM XML Schema, please see Section 3.5 of FEA DRM, which is included in Appendix D.

3.4.3.2 Context

The State should consider setup of multiple taxonomies for Data Context applying DRM XML Schema, as that will support how the data is searched. The Data Context is to be registered with California Metadata Registry.

For a more detailed explanation regarding the concepts and standards pertaining to respective attributes specified in DRM XML Schema, please see Section 4.5 of FEA DRM, which is included in Appendix D.

3.4.3.3 Sharing

We recommend defining Data Exchange and Data Access Services as metadata applying DRM XML Schema. The metadata then is to be registered with California Metadata Registry, which in turn would enable its discovery.

For a more detailed explanation regarding the concepts and standards pertaining to respective attributes specified in DRM XML Schema, please see Section 5.5 of FEA DRM, which is included in Appendix D.

3.4.4 Service Orientation

Service orientation is a design paradigm where all work and data are abstracted behind a well defined 'service', the architectural principles behind Service Oriented Architecture (SOA). Service Oriented Integration (SOI) is an approach to defining integration architectures based on the concept of a *service*. It is a subset of SOA relating to web services and application integration. It applies successful concepts by Object Oriented development, Component Based Design, and Enterprise Application Integration technology. The goal of SOI can be described as bringing the benefits of loose coupling and encapsulation to integration at an enterprise level.

3.4.4.1 Service Oriented Integration

In order to describe SOI, it is first necessary to define what we understand by a "service" in this context. The most commonly agreed-on aspects of the definition of a service in SOI are:

- It is defined by an explicit, implementation-independent interface
- It encapsulates a reusable business function
- It is loosely bound and invoked through communication protocols that stress location transparency, interoperability, and security.

By explicitly defining an interface, we can hide the specifics of implementation for a business process or function. In addition, aspects of the system such as platform it is based on are hidden from service consumers, thus providing flexibility for change to the platform without affecting the consuming application. The use of interfaces to define and mediate various aspects of service interactions is discussed in Section 3.4.4.1.2 - *Aspects of Service Interactions*.

After the function has been encapsulated and defined as a service in an SOI, it can be used and reused by one or more systems that participate in the architecture. The encapsulation of services by interfaces and their invocation through location-transparent, interoperable protocols are the basic means by which SOI enables increased flexibility and reusability. In the following sub-sections we will cover:

- Reasons behind recommending SOI
- Aspects of service interactions
- The concept of choreography
- An implementation strategy for SOI
- The concept of an enterprise service bus
- Web services and how they relate to SOI
- Leveraging third party XML schemas

3.4.4.1.1 Why SOI?

Using SOI is recommended as a part of the strategy for the following reasons:

- There are a multitude of technologies and platforms used statewide.
- Business processes are complex and can be decomposed into interactions between humans and systems or systems and systems.
- A loosely coupled approach minimizes the “ripple effect” on other interfacing systems as changes are made to one system.
- The traditional approach of using a single integration solution statewide would be expensive and time consuming.
- A single integration solution can’t support both internal and external partners.
- Not all integration technologies work as well across a wide area network or the Internet as they do across a local area network.

3.4.4.1.2 Aspects of Service Interactions

A basic tenet of SOI is that services are *loosely coupled*. By loosely coupled services, we mean that client of a service is essentially independent of the service. Both requestor and provider have a minimal knowledge of each others’ code in terms of programming language and platforms they are executing on. All services implemented in practice have either a coupled or a de-coupled aspect of service interactions. Change made to any aspect of a service that is *coupled* requires a subsequent change to application code of the requester or the provider or, as in most cases, to both. For example, the business behavior (the function and data model) obviously must be coupled in order for the requester and provider to interact. If a change is made by the requester or the provider to any aspect of a service that is *decoupled*, then there should be no need to make subsequent changes in the other parties.

Furthermore, coupled and decoupled are not the only two relationships that can exist for an aspect of a service between the requester and the provider. The interactions between requester and provider must also be secured, and the relationship between their transactional models will have to be understood in order to define how failures will be handled.

3.4.4.1.3 Choreography

We can classify SOI services that correspond to business processes in the State of California as “fine-grained” and “large-grained”. By “fine-grained” we mean a reusable service that is mapped to a single data asset or entity. A "large-grained" service is defined as a reusable service that choreographs[9] the execution of multiple fine-grained services. At either level of granularity, it is important that a service definition encapsulates function well enough that it is reusable.

3.4.4.1.4 Implementation

The encapsulations of reusable business function, the achievement of loose coupling, and the definition of appropriate levels of granularity are business issues as much as technology issues. It’s not an easy task to grasp these difficult principles. Thus SOI cannot be successful without skilled architects and designers who understand these principles and are able to articulate them.

Implementation of SOI and associated infrastructure is a long-term endeavor for the State. It is a serious long term commitment and involves all of the usual hard business decisions, questions of data, process ownership, and costs that are typical for any integration project. Adopting a SOI approach has advantages such as:

- Support for existing technologies as technologies are ruled in or ruled out.
- Support for legacy implementations such as mainframes is possible.

In order to implement an SOI, both applications and infrastructure must support the SOI principles of implementation independence and loose coupling. Applications can be enabled by the creation of service interfaces to business functions, either directly or through the use of adapters. To enable the infrastructure at the most basic level, an Enterprise Service Bus (ESB) is recommended. It enables the provision of capability to route and transport service requests to the correct service provider.

3.4.4.1.5 Enterprise Service Bus

An ESB facilitates SOI implementation. An ESB is a software infrastructure that simplifies the integration and flexible reuse of business components within a SOI. An ESB provides a dependable and scalable infrastructure that connects disparate applications and IT resources, mediates their incompatibilities, orchestrates their interactions, and makes them broadly available as services for additional uses. Some of the considerations towards selecting an ESB are:

[9] In this context, a ***choreograph*** is similar to a workflow where multiple services are executed in sequence to do a larger unit of work.

- **Performance and Scalability** - It is essential that the infrastructure used for the state provide performance and scalability. It is critical that it not create bottlenecks or limit the throughput of data it can carry to and from connected resources.
- **Security, Reliability, and Availability** - The state's security, reliability and availability standards are stringent. The ESB selected should be able to meet those standards.
- **Distribution -** In an SOI, services and service orchestrations will interact with services spread across an agency, and between agencies. An ESB provides the communications facilities which link agencies together with services and messaging support. Both queuing and publish-and-subscribe behavior are supported within the ESB.
- **Flexibility -** The ESB should enable flexibility towards change in orchestration, rules, data mapping, and relationships between applications with minimal effort and disruption.
- **Visibility and Control** - The ESB should manage and monitor the infrastructure as well as the processes and services deployed within it. Additionally, the ESB should make it easy to deploy and upgrade services remotely from a central location.

The true value of the Enterprise Service Bus (ESB) concept is to enable the infrastructure for SOI in a way that reflects the needs of today's enterprise: to provide suitable service levels and manageability, and to operate and integrate in a heterogeneous environment. The implications of these requirements go beyond basic routing and transport capability, and they are described in the standards section of Section 8. The ESB should enable the substitution of one service implementation by another with no effect to the clients of that service. This requires the service interfaces that are specified by both the SOI and the ESB allow client code to invoke services in a manner that is independent of the service location and communication protocol that is involved.

Why ESB?

The reasons for recommending ESB in SOI are:

- The ESB supports multiple integration paradigms. It supports Service Oriented Architectures, message-driven architectures and event-driven architectures.
- The ESB centralizes control and distributes processing.
- The ESB enables wider connectivity to legacy systems.
- The ESB provides choreography of services

3.4.4.1.6 Web services

Web services are a set of technology specifications that leverage existing proven open standards such as XML, URL, and HTTP(S) to provide a new system-to-system communication standard. Based on this communication model, additional higher-level Web services standards have also been defined to address transactions, security, and business processes.

The Web revolutionized how people interact with systems. Web services similarly bring a revolution towards system to system interactions. By adopting Web services, the cost can be dramatically reduced for both inter- and intra-business interactions while achieving a higher level of efficiency. Basic Web services define interactions among Service Requesters, Service Providers, and Service Directories as follows:

1. Service Requesters find Web services in a Universal Description Discovery and Integration (UDDI) Service Directory[ix].
2. They retrieve Web Service Definition Language[x] (WSDL) descriptions of Web services offered by Service Providers, who previously published those descriptions to the Service Directory.
3. After the WSDL has been retrieved, the Service Requester binds to the Service Provider by invoking the service through SOAP.
4. The basic Web services are often described in terms of SOAP, WSDL, and UDDI, each of which we define and discuss.

However, it should be noted that each of these standards can be used in isolation, and there are many successful implementations of SOAP alone, or SOAP and WSDL, in particular.

Messaging – SOAP

SOAP is an XML messaging protocol standard from World Wide Web Consortium (W3C). SOAP defines a framework within which messages contain headers and a message body. It is independent of any specific transport protocol. In practice, SOAP is most often communicated over HTTP or HTTPS. SOAP makes no reference to characteristics of interactions such as security and transactions. Since SOAP headers support extensibility, these aspects are being added to the Web services specifications as extensibility elements. The use of SOAP over specific protocols, such as HTTP, is usually documented as SOAP/HTTP, SOAP/JMS.

Description – Web Services Definition Language (WSDL) [x]

"WSDL is an XML format standard from W3C for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. The operations and messages are described abstractly, and then bound to a concrete network protocol and message format to define an endpoint. Related concrete endpoints are combined into abstract endpoints (services). WSDL is extensible to allow description of endpoints and their messages regardless of what message formats or network protocols are used to communicate"[10].

Like SOAP headers, the WSDL specification is extensible and supports the additional aspects of a service interaction, such as security and transactions.

[10] Taken from the W3C specification for WSDL. WSDL is an open standard and is copyrighted in 2001 by Ariba, International Business Machines Corporation, and Microsoft. All Rights Reserved. http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231.

Discovery – UDDI [ix]

"Universal Description Discovery & Integration (UDDI) is the definition of a set of services supporting the description and discovery of (1) businesses, organizations, and other Web services providers, (2) the Web services they make available, and (3) the technical interfaces which may be used to access those services. Based on a common set of industry standards, including HTTP, XML, XML Schema, and SOAP, UDDI provides an interoperable, foundational infrastructure for a Web services-based software environment for both publicly available services and services only exposed internally within an organization"[11]. The original UDDI classification was based on U.S. government taxonomy of businesses, and recent versions of the UDDI specification have added support for custom taxonomies.

A public UDDI directory is provided by IBM, Microsoft, and SAP, each of whom runs a mirror of the same directory of public services. However, there are many patterns of use that involve private registries; see Steve Graham's articles[xi, xii].

Web Services Interoperability

A unique feature of Web services is that it is a relatively high-level set of integration protocols with near-ubiquitous support in the IT industry. Several projects have used, and more are continuing to use, Web services standards to perform integrations between different platforms.

In order to facilitate the development of truly interoperable Web services standards, the Web Services Interoperability Organization (often referred to as the WS-I) was formed in February 2002. The WS-I aims to promote interoperability of Web services implementations by publishing *profiles*[12], which are descriptions of conventions and practices for the use of specific combinations of Web services standards through which systems can interact. WS-I published WS-I Basic Profile[13] in February 2007.
Though interoperability can be achieved using Web services where WS-I profiles do not exist, it is recommended to leverage WS-I Basic Profile to ensure proper interoperability.

Web Services Security[14]

Security is a key aspect of a good Web services architecture. Given the criticality of the shared information, it is paramount to support a comprehensive security model for the system. Some of security requirements for the system are authentication, authorization, privacy, trust, integrity, confidentiality, secure communications channels, federation, delegation, and auditing. We propose usage of Security Token Service (STS) in conjunction with Web services security standards as the model to address the requirements. STS is defined in WS-Trust specifications. STS brings together security

[11] Taken from the OASIS UDDI Specification - Copyright © 2000 - 2002 by Accenture, Ariba, Inc., Commerce One, Inc. Fujitsu Limited, Hewlett-Packard Company, i2 Technologies, Inc., Intel Corporation, International Business Machines Corporation, Microsoft Corporation, Oracle Corporation, SAP AG, Sun Microsystems, Inc., and VeriSign, Inc. All Rights Reserved."

[12] Web Services Interoperability Organization published profiles - http://www.ws-i.org/deliverables/matrix.aspx

[13] Web Services Interoperability Organization specified a set of non-proprietary set of web service specifications. http://www.ws-i.org/Profiles/BasicProfile-2_0(WGD).html

[14] Copyright © OASIS® 1993–2007. All Rights Reserved. OASIS trademark, IPR and other policies apply.

technologies such as Public Key Infrastructure (PKI) and Kerberos. The following Web services security specifications meet the other requirements:

- **WS-Security:** a standard set of extensions to SOAP messages that could be used to ensure content integrity and confidentiality. It is flexible and works with a variety of security models such as X.509 certificates and Kerberos tickets.
- **WS-Policy:** a framework that provides capabilities and constraints of the security policies on components such as intermediaries and endpoints.
- **WS-Trust:** a framework that enables Web services to securely interoperate.
- **WS-Privacy:** a standard to enable communication of privacy policies within a system.
- **WS-SecureConversation:** a standard defined for exchanging security information between Web services.
- **WS-Federation:** describes standards for achieving level of trust between disparate systems in a federated environment.
- **WS-Authorization:** describes a standard for managing authorization and access policies of data.

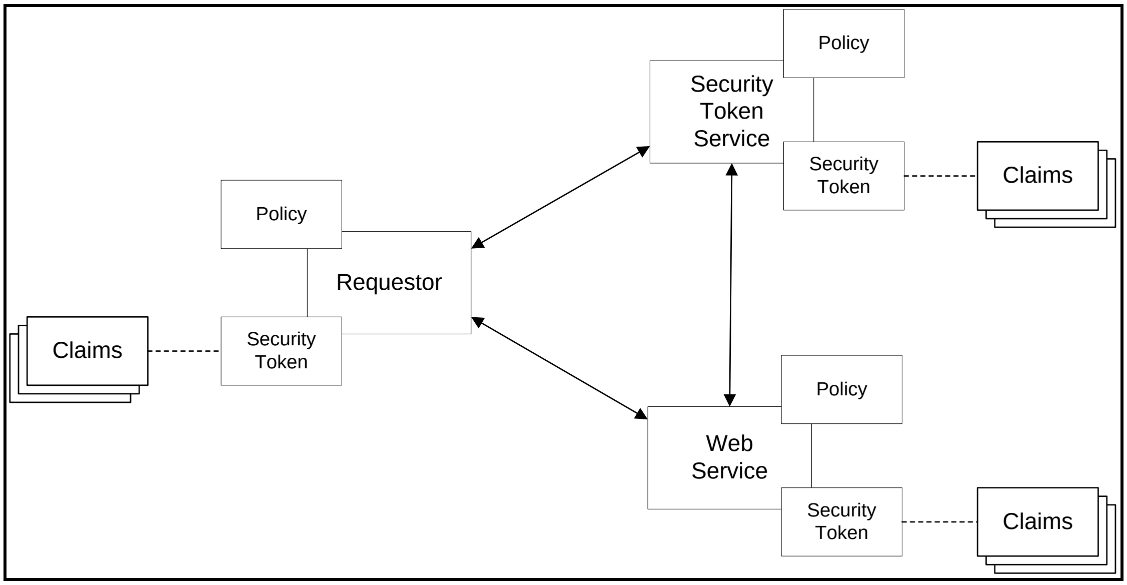


Figure 3-17 – Security Token Service[xiii]

The process followed in the model[xiii] is:

- As specified in the policy, a Web service can require that an incoming message prove a set of *claims* (e.g., name, key, permission, capability, etc.). The service may ignore or reject a message that arrives without the required claims.

- Security tokens are associated with messages and sent by a requester. When the Web service verifies the claim, it processes the message.
- In addition, a *security token service* can provide the necessary claims though a web service to the message sender. These security token services may in turn require their own set of claims.

The general messaging model of using claims, policies, and security tokens supports any security capability.

For public internet applications though, the ability to withstand concerted denial-of-service attacks is a higher priority. The security requirements may be combined in many ways and specified at many different levels. A successful approach to Web service security requires a set of flexible, interoperable security primitives that through policy and configuration enable a variety of secure solutions. Listed below are sample security scenarios that should be considered when setting up Web Services security:

- **Direct Trust using Username/Password and Transport-Level Security -** In this scenario, a Web service authenticates a requester using a username and password with transport security.
- **Direct Trust using Security Tokens -** In this scenario, a requester establishes direct trust using X.509 certification and Kerberos service tickets (ST).
- **Security Token Acquisition -** In this scenario, a Web service authenticates a requester using a security token stored independently from the message.
- **Firewall Processing -** In this scenario, firewalls leverage a security model for greater degrees of control.
- **Issued Security Token -** In this scenario, security tokens issued by certification authorities are used for basic authentication.
- **Enforcing Business Policy -** In this scenario, security tokens issued for a business process are used.
- **Privacy -** In this scenario, clients and services can communicate their privacy policies.
- **Web Clients -** In this scenario, a Web browser is the client requesting information.
- **Mobile Clients -** In this scenario, mobile clients can securely interact with Web services.
- **Access Control -** In this scenario, Web services security supports traditional access control list-based security.
- **Auditing -** In this scenario, auditing is used to track security-related activities and incidents.

One of the benefits with SOI, Web services and XML is the large number of standard XML schemas that are being created at both the federal space and the commercial sector. These schemas provide a common definition for standard data that is shared within government, medicine, and business.

3.4.4.1.7 Leveraging Third Party Schemas

Software solutions can be categorized in three types of approaches as follows:

- **Framework -** A software framework is typically sponsored by a standards committee (e.g., OASIS, W3C, ISO) and provides standard and process definitions and is often supplemented by little software. Generally the preexisting code is basic in nature and needs to be heavily adapted.

  Pros: High flexibility enables delivery to a broad range of requirements; development can be done by general developers (e.g., in-house).

  Cons: Depending on the preexisting code base, everything from basic functionality to customizations needs to be developed.

- **Packaged Solutions -** Vendors within the software industry advertise their software packages as solutions for specific situations. Product comparison and vendor research is required to determine the maturity level of the software and customer feedback.

  Pros: Fulfills the majority (e.g., 80%) of customer needs.

  Cons: Customizations are typically difficult as vendor focus was on providing a product not a framework (meaning obtaining the remaining 20% requires specialized developers often resulting in non-reusable code).

- **Turnkey Solutions -** Typically large vendors known as a capacity within their field offer end-to-end software solutions.

  Pros: Most customer needs can be satisfied; Cost easier aligned with budget (prioritization based on available funding).

  Cons: Vendor lock-in; software development only through highly trained specialists (typically through the same or affiliate vendor); often large development cost for uncommon customizations.

A number of third party schemas were evaluated for the report. Please see Appendix F for the list and findings.

3.4.4.2 Considerations

To support service orientation, we recommend a SOI approach. To facilitate SOI implementation we recommend usage of an ESB. An ESB will provide a dependable and scalable infrastructure in addition to the support of Web services standards. Web services standards that we recommend are SOAP, WSDL, UDDI, WS-Security, WS-Trust, WS-Policy and WS-Choreography.

Finally, to enable a comprehensive security architecture for the system, we recommend usage of Security Token Service (STS) in conjunction with the Web services security standards.

3.4.5 Trading Partner Framework

Trading partners are organizations which agreed to conduct business collaborations by exchanging business transactions with each other. This agreement is typically formalized in a Trading Partner Agreement (TPA), which details roles and responsibilities for business activities. Trading partners are typically assigned a unique identifier which allows them to be recognized within the automated business collaboration.

Delivery channels describe communication capabilities to exchange business transactions. Trading partners agree on an exchange mechanism and a transport protocol to transfer messages securely and in the proper sequence. The document exchange protocol defines how business documents are received, encrypted (incl. application of digital signatures), and handed over to the transport protocol which is responsible for transmission to the other trading partner. The transport protocol handles message delivery (incl. receipt acknowledgement) using communication protocols such as (S)HTTP, SMTP, or (S)FTP as well as the transport security.

Business activities are conducted between the roles authorized to participate in a collaboration, and can either be a business transaction or a business collaboration.

A business transaction is a unit of work conducted by two or more trading partners, one with the initiator (or From) role, and the others with the responding (or To) role. A business transaction contains business data and generates a business signal with an agreed format, sequence, and time period resulting in a definite state of success or failure.

Business transactions are validated by the application layer. If any of the rules identified in the agreement is violated, the transaction is terminated and the initiating partner is informed. Some examples for this are:

- trading partner not recognized within the collaboration network (e.g., incorrect party id)
- requestor not authorized for business transaction
- message format not recognized (e.g., invalid XML, incorrect encryption)
- response to a non-existing request

Business collaboration leverages defined roles within predefined business transactions, such as buyer and seller, and spans one or more business transactions to achieve a specified outcome. For example a requesting action such as a purchase order request is followed with a responding action such as a purchase order acceptance.

3.4.5.1 Considerations

The State should consider developing a standard TPA to be used for each trading partner who wishes to participate in a collaboration or information exchange. The standard TPA should also detail the on-boarding process and expectations of how new trading partners will join the collaboration.

In addition, a separate addendum to the standard trading partner agreement should be developed to cover all nonstandard data needs. This will provide a common contractual basis across all trading partners while allowing more flexibility with specific trading partners on an as needed basis.

Since external partners will be supported, the opportunity for a security breach is much higher than internal user communication. Therefore, for the infrastructure approach, we recommend creating a data management zone (DMZ) to establish an enterprise boundary to separate external users from internal users. We further recommend usage of two separate firewalls and reverse proxy servers, as they are a de-facto industry standard, which also makes them a cost effective security measure.

3.4.5.2 Trading Partner Agreement

A TPA is a written contract between trading partners that sets expectations related to the exchange of information and specifies technical details on how electronic transactions are conducted.

For example, a TPA may outline policies and processes for information exchange, list duties and responsibilities for each party, spell out permitted electronic transactions and their technical details, and specify service level agreements for process flows (such as time of day, quantity, quality, and possible consequences of not meeting expectations).

3.4.5.2.1 Content

While a trading partner agreement is a legally binding contract, there is no single way of drafting such an agreement. In many cases, an organization will determine what should go into its TPA, unless the organization committed to the standards of a larger information exchange network.

TPAs cover a variety of aspects around information exchange, such as:

- **Keys & Identifiers -** Some identifiers cannot be looked up within a database and need to be defined. For example if information is exchanged between multiple trading partners, each trading partner is required to have a unique identifier which identifies it among all trading partners. Unless a standard registry is utilized, each trading partner needs to be assigned a unique identifier.
- **Format & Resources -** This section of the TPA contains technical details about the electronic data exchange, such as:
    - file formats
    - definition of delimiters
    - string length limitations
    - minimum and maximum acceptable file sizes
    - domain names / IP addresses and port numbers
    - directory paths

  - any other details related to the information exchanged

- **Processing Requirements -** Specifies how much information of what type is exchanged during which time periods. For example, sending too little information may not meet the needs of the business while sending too much information may have a technology impact. This information is used to size the infrastructure accordingly so that the information can be processed while meeting any service level agreements that are in place.

- **Transaction Choreography -** Some transactions may require more complex workflows such as acknowledgement of receipt or automated handling of exceptions. In this section, the sequencing of electronic documents between several trading partners is specified for each transaction type. For example, a purchase order is submitted, followed by a confirmation of receipt, then an invoice is sent, followed by a confirmation of receipt, then a shipping statement is sent.

- **Financial Arrangements -** This section details the business portion of a TPA, including which partner covers what technical costs, charges, or fees that may be due, and potential penalties for not meeting service level agreements. This section ties into the service level agreements that are in place, if any. If federal funds are involved then the federal government may be involved in the agreement as well.

- **Security -** This section details expectations and requirements for encryption, electronic signatures, Public Key Infrastructure (PKI), and equivalent mechanisms. This section may also specify what data is acceptable to be electronically transmitted and what data may not be exchanged.

- **Taxonomy & Codes -** Extensive lists may be used to map key codes from one trading partner system to another trading partner system, essentially translating terms between application software and making the information understandable. For example the term “jacket” from one system needs to be translated into “coat” to be processed by another system.

- **Exception Handling -** This section of the TPA covers so called human intervention, expectations of each trading partner on how to handle situations where the automated system encountered issues. Here individuals are assigned roles and responsibilities and timelines for restoration are set. Some examples of exception handling are: the transaction may be routed to a work queue on an exception or may be rejected outright. This section may also cover how trading partners will address preventive action for future events.

- **Testing -** Most information exchanges have an on-boarding phase, where the systems are setup before business as usual can begin. This phase requires hands-on work on both side and coordination. This section of the TPA outlines how testing is conducted, how test transactions are distinguished from real business transactions, technical details and timelines for initial data loading, and what conditions are to be met before going live.

- **Contact Info & Escalation -** This section of the TPA contains the trading partner names with telephone numbers and escalation paths for customer service and support assistance.

3.4.5.2.2 Change Management

The State should develop separate, supplemental documents to capture changes or additions to an existing trading partner agreement. Many times an organization will choose to amend the original agreement rather than a complete rewrite to avoid a re-negotiation of the main agreement.

It is best practice to develop a standard form suitable for all trading partners and establish separate, supplemental documents for all nonstandard data needs for specific trading partners. This will provide a common contractual basis across all trading partners while allowing more flexibility with specific trading partners on an as needed basis.

3.4.5.3 Data Management Zone

A Data Management Zone (DMZ), occasionally called a demilitarized zone, is a customer facing network separate from the internal local area network (LAN). It is used to expose internal services to a non-trusted network such as the Internet.

The DMZ is surrounded by two separate firewall systems (front and back end) and contains lightweight servers which act as proxies for internal services. External users access services over an incoming broadband connection through the Internet facing (front end) firewall which makes these services available via a public IP address.

The proxy servers within the DMZ are registered with the corporate facing (back end) firewall where they are allowed to relay requests to back end systems over the internal LAN using an internal IP address not available to the public.

The advantage is that all hosts within the DMZ are managed by the enterprise and trusted by the corporate facing (back end) firewall. Potential attackers are not given a direct access path to the corporate network and are limited to interaction with the exposed services. The DMZ establishes an effective enterprise boundary by separating external users from internal users.

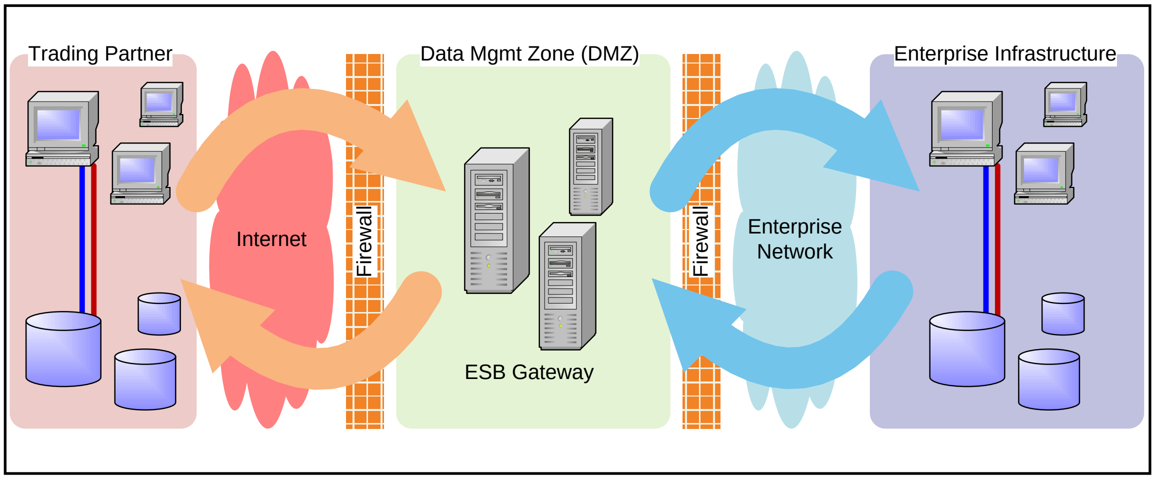


Figure 3-18 – Trading Partners Network

Typical services setup in a DMZ:

- **Web Servers -** Web servers act as proxies to communicate to internal database systems. For increased security, an application server should be used as a medium for communication between the web server and the database server.
- **E-mail Servers -** The mail server in the DMZ acts as a proxy for the internal mail server. Incoming mail is passed to the internal mail server and the internal mail server passes outgoing mail to the external mail server.
- **Proxy Servers -** Internal users should not be allowed to bypass the DMZ defenses by accessing the Internet directly. Proxy servers provide an easy means of monitoring user activities to make sure that no confidential (or illegal) content gets in or out of the enterprise.
- **Reverse Proxy Servers -** A reverse proxy server provides indirect access to internal resources from an external network (such as back office application access for employees). Usually, application layer firewalls are utilized, as they are able to monitor the content of the traffic instead of being limited to port numbers only like a packet filter firewall.

3.4.6 Consolidated Data Warehouse

One of the major design considerations and requirements is to support ‘single source of truth’ reporting for the Governor’s office, Legislature, and decision makers within the State government. Ensuring the same information is used to address questions is critical. Master data that is created in the Shared Data Space has been harmonized, data ownership has been identified, and the update business rules have been identified. However, the data may not be in a form that is easily reported on. That is why a dimensional data store may be necessary to support future reporting requirements. Figure 3-18, illustrates where both a data mart and a data warehouse fit into the overall solution.

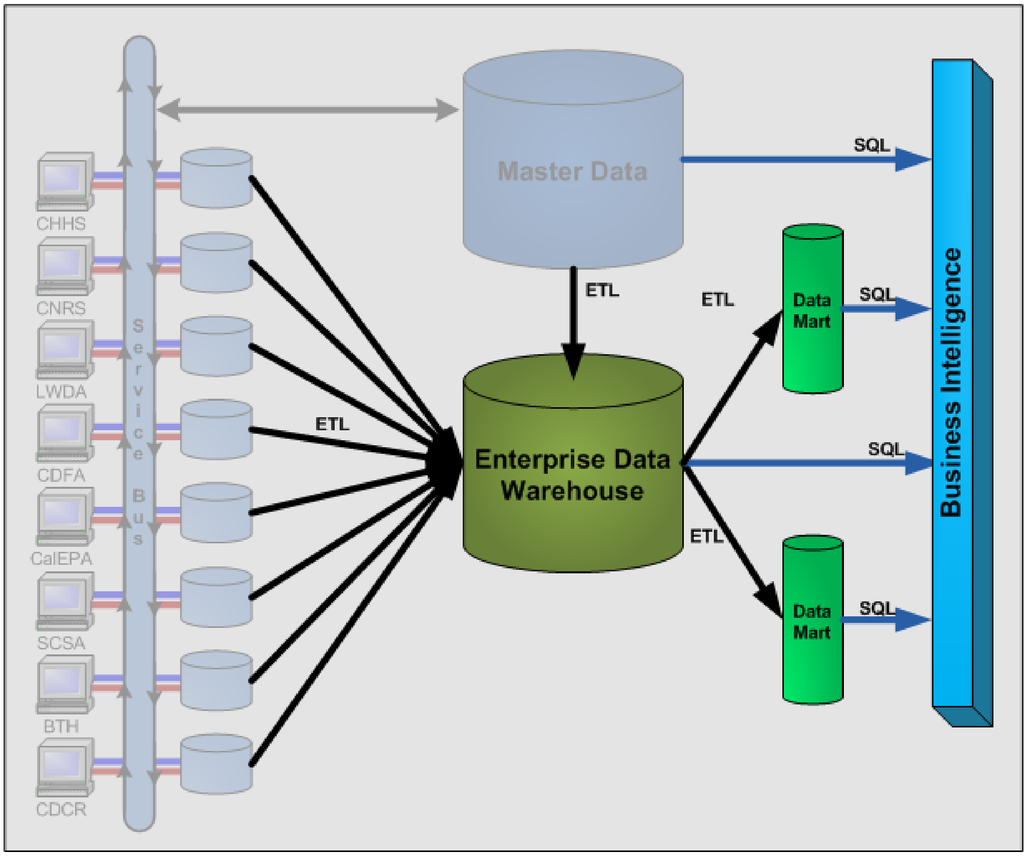


Figure 3-19 – Consolidated Warehousing

The ETL feeds in Figure 3-18 coming directly from the agencies are included to illustrate that additional data, that is not yet available through the master data repository can be added to the data warehouse. It should be noted, that there is a cost to this approach in that the data has not been de-duplication nor has it been validated and it represents the view from only one agency. Therefore, care should be given to minimize the use of these one-off interfaces.

3.4.6.1 Consolidation Opportunities

The goal of the State is to improve efficiency and the quality of services provided to its constituents while reducing cost. To realize these improvements we recommend centralizing and consolidating the State's data warehousing repositories. The focus is on reducing this number to not only improve the quality of the data, but to also have a consistent analysis from the data, leading to more efficient decision making and better decisions. Multiple data repositories increase the risk of inadvertent use of stale and/or inaccurate data, thus giving rise to erroneous conclusions. In addition, maintaining, synchronizing, and troubleshooting a data issue across multiple reporting databases is costly.

In support of reducing the number of data warehousing repositories across the State, information from the big eight agencies was gathered detailing their reporting databases.

The following data usage pattern was observed and Organization, Person, and Location, all have about equal value to the business.

Repository Count	Repository Data
145	Organization
136	Person
84	Location
69	Financial
42	Reference
32	Resource
23	License
16	Unstructured
15	Medical Records
14	Services
14	IT
10	Facility
8	Law Enforcement
8	Unknown
3	Regulations
1	Government

Table 3-1 – Data Subject Area Count by Reporting Repository

As information is normalized in to the master data repository, the opportunity for generating a reporting repository (e.g., Data Mart) exists. If additional data is needed, it can be sourced in one of two ways. It can either be instantiated in the master data repository or provided to the reporting data base when needed from an ETL process.

3.4.6.2 Advantages of Data Marts

There are some real advantages with building data marts over a full data warehouse. For highly complex enterprises, which the State of California clearly is, a data mart is usually the simplest initial approach. This is not an 'either or' decision, and many mature organizations support both data marts and an enterprise data warehouse. A few advantages for starting with data marts are:

- **Flexibility -** Data from the data warehouse can be restructured to seamlessly feed into downstream applications or systems (e.g., third party online analytical processing (OLAP) applications). This strategy can also be leveraged as a proving ground to demonstrate viability and return on investment (ROI) of an application prior to migrating it to the data warehouse.
- **Performance -** Frequently used subsets of a data warehouse can be moved to separate hardware, improving end-user response time, assuming network latency is not a performance bottleneck. In the case that network latency is an issue such hardware can be located on a local network, offering higher network bandwidth.

- **Cost Efficiency -** Data marts can be efficiently recreated (possibly on a scheduled basis), as they are duplicated subsets of a data warehouse. Data retention requirements become minimal, as the data is easily replaceable. This recreation can also reduce overall hardware cost and minimize new system requirements, as opposed to the cost incurred for implementing a full data warehouse.
- **Security -** Data can be separated by data context for confidentiality. For example data classified as limited or restricted can be isolated from public data.
- **Low Maintenance -** Data marts have more clearly defined usage patterns compared to a full data warehouse, making troubleshooting and maintenance less demanding on existing personnel.

3.4.6.3 Considerations

The recommended approach is to initially leverage data marts over a large data warehousing initiative. Since the agencies are currently supporting their own warehousing initiatives, there is little downside on taking this incremental approach. Since data marts are small disposable reporting repositories, they can be easily created and destroyed to meet the reporting needs of the business. As the requirements for a more persistent dimensional data store emerge, the emphasis will increase on using an enterprise data warehouse. The Data Governance committee will evaluate the reporting needs of the business with the shared data that is available. As more data is shared, more and more reporting can be centralized.

Since an enterprise data warehousing initiative is not trivial, a special project phase will be necessary to gather the requirements for the warehouse. Enough data will need to be available or extracted to support these requirements. This approach is not evolutionary in that there is a minimal set of data needed within a data warehouse to make the approach viable. Therefore, initially leveraging data marts for target reporting is the best way to get started.

3.4.7 Geospatial information

A more complete write-up of geospatial data is provided in the California Geospatial Framework Draft Data Plan[15]. *"Geospatial information is defined as data pertaining to the geographic location and characteristics of natural or constructed features and boundaries on, above, or below the earth's surface; esp. referring to data that is geographic and spatial in nature. Geospatial is a term used to describe both spatial software and analytical methods with geographic or terrestrial datasets."* Webster's New Millennium Dictionary.

In reference to the context of data sharing, geospatial data provides the geographic attributes or context of a location. Geospatial data, however, tends to include very large landscapes and can contain extremely detailed information. Geospatial data tends to have two primary structures; 1) Raster and 2) Vector. Raster data is data of pixels, often millions of pixels, where each pixel contains a single value or even a stack of values. Raster data is commonly composed of aerial or satellite imagery pictures. Vector data includes points, lines, or areas that describe a location using these abstract shapes. For instance, point data often represents houses, trees or single point locations. Line data often represents streams or roads. Area data might describe boundaries or landscape features like natural vegetation.

The framework data themes have been developed by the federal government's National Spatial Data Infrastructure (NSDI). These themes are being extended by the State of California. These themes are layers of data that will be maintained in the GIS repository. Figure 3-21 illustrates the draft geospatial framework that the State of California is proposing. Several examples on how the GIS information can be accessed are:

- **X –** The X coordinate on a Cartesian coordinate system using the California defined standard projection and datum.
- **Y** - The Y coordinate on a Cartesian coordinate system using the California defined standard projection and datum.
- **Datum *(Defaulted)*** - "A datum is a geodetic reference system that specifies the size and shape of the earth, and the base point from which the latitude and longitude of all other points on the earth's surface are referenced". - *Canada's National Statistical Agency.* The default is North American Datum of 1983 (NAD83) and World Geodetic System of 1984.
- **Projection *(Defaulted)* -** "To represent a curved surface such as the Earth in two dimensions, you must geometrically transform (literally, and in the mathematical sense, "map") that surface to a plane". *The MathWorks, Inc.* The default projection is still being decided at the time of the writing of this strategy.

Or

- **x –** The x coordinate of address location within the US National Grid Coordinate system.

[15] Several outreach workshops have been conducted by Michael Baker Jr., Inc. across California. The company is authoring the California Geospatial Framework Data Draft Plan (the California Geospatial Framework Plan). More information can be found at http://www.cgia.org/geospatial-draftplan.htm

- **y –** The y coordinate of address location within the US National Grid Coordinate system.
- **USNG –** US National Grid – is a simplified plane coordinate system that goes across jurisdictional boundaries and map scales

Or

- **Latitude** – "the angular distance north or south from the equator of a point on the earth's surface, measured on the meridian of the point". *Webster's New Millennium Dictionary*
- **Longitude** – "angular distance east or west on the earth's surface, measured by the angle contained between the meridian of a particular place and some prime meridian, as that of Greenwich, England, and expressed either in degrees or by some corresponding difference in time". - Webster's New Millennium Dictionary
- **Datum *(Defaulted)*** - "A datum is a geodetic reference system that specifies the size and shape of the earth, and the base point from which the latitude and longitude of all other points on the earth's surface are referenced". - *Canada's National Statistical Agency.* The default is North American Datum of 1983 (NAD83) and World Geodetic System of 1984 (WGS84).

There are many possible approaches to linking physical addresses to the GIS data, but there are only three discussed in the strategy. Additional requirements gathering will need to be conducted to determine the final approach.

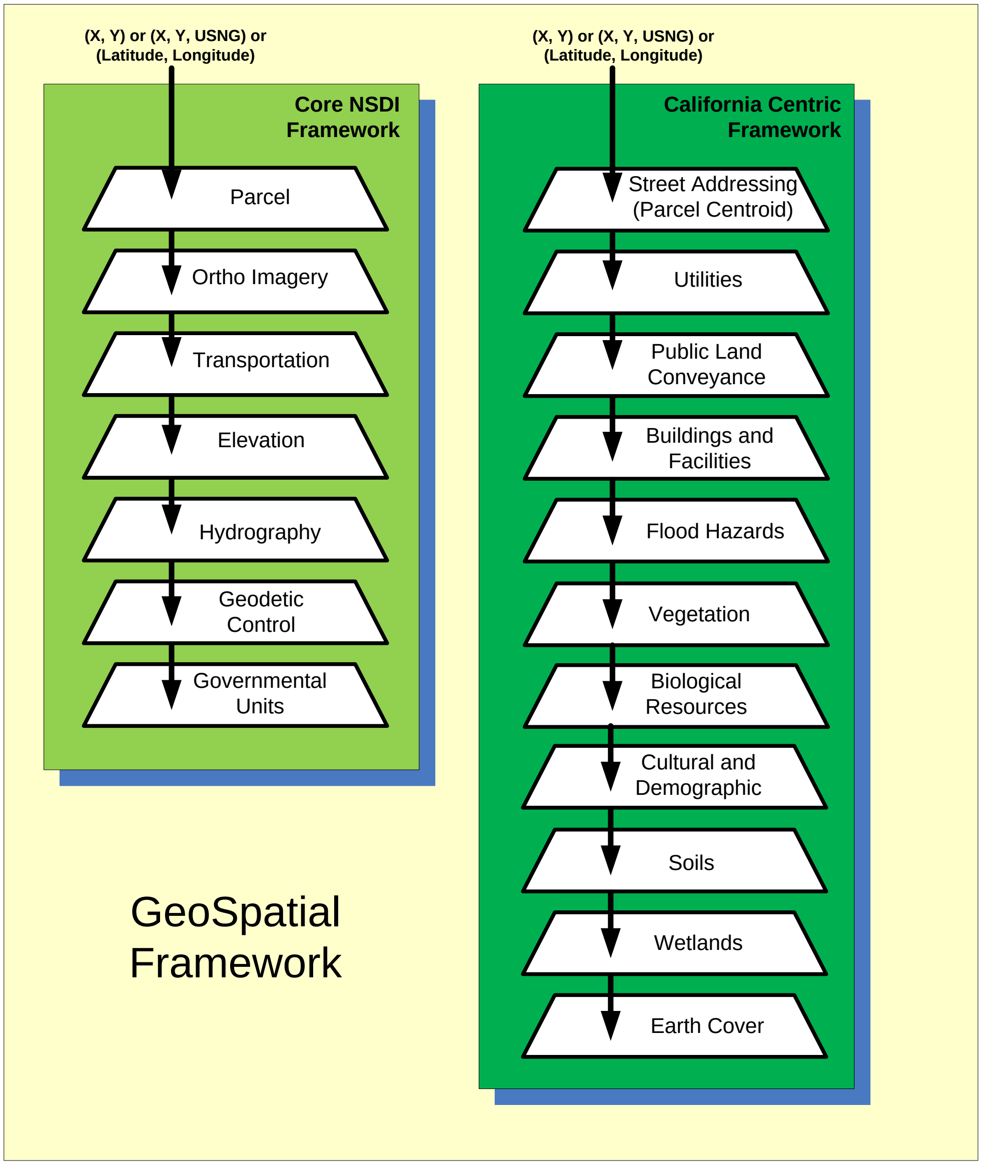


Figure 3-20 – California Geospatial Framework

The simplest approach to linking a physical address to the GIS information is to use a point (i.e., X and Y), the California default **datum** (NAD83), and the California default **projection**[16]. Another approach to consider is the same approach suggested by the FGDC Street Address Standard Working Draft. Once a point is defined (i.e., X and Y) and US National Grid Coordinate, a point in space is identified with a high level of accuracy. Finally this relationship can be built using the Latitude and Longitude and defaulting to the **datum** to the North American Datum of 1983 (NAD83) or the World Geodetic System of 1984 (WGS84).

From these points in space we can determine its relative relationships with other points, lines, and polygons in space. In Figure 3-21, eighteen different "layers" of data are

[16] The State of California has not yet chosen a default projection model but one is currently under consideration.

illustrated. As an illustration, identifying a point in space will allow identification of not only, for instance, the parcel, but also the governmental units that contain the point.

Alternatively, with a formatted address, the center point of the parcel can be found. The center of a parcel is information that is captured with each parcel and defines a point that can be queried against the other layers. Now the address will have little usefulness with some of the layers (e.g., streams) since a physical address does not exist, however, in some situations accessing the information in this way becomes extremely useful.

To tie this all together, once an address is identified, providing a reference point, the geospatial information relating to that address can also be selected. The opposite can also be true. Consider the situation where the State wants to know the relative position of homes with respect to a flood plain. By traversing the relationship the other direction, this information can be identified.

3.4.8 Minimizing Impact to Business

Each agency is autonomous in their charter and is performing well within their core business. Due to this autonomy, the size of the agencies and the overall complexity of their IT environment, it is desirable to provide a smooth transition into data sharing. To completely rework all of an agency's applications to support data sharing is overwhelming, expensive, and risky. The goal is to add value to the business without negatively impacting the current business processes. This goal can be accomplished by treating data as a service.

The stove-piped data is still maintained in the agency, but only the data that needs to be shared is a candidate to be maintained as a shared asset. All data is not necessarily created equal. Some data has high value to the enterprise as a whole, while other data may only be valuable to a single department.

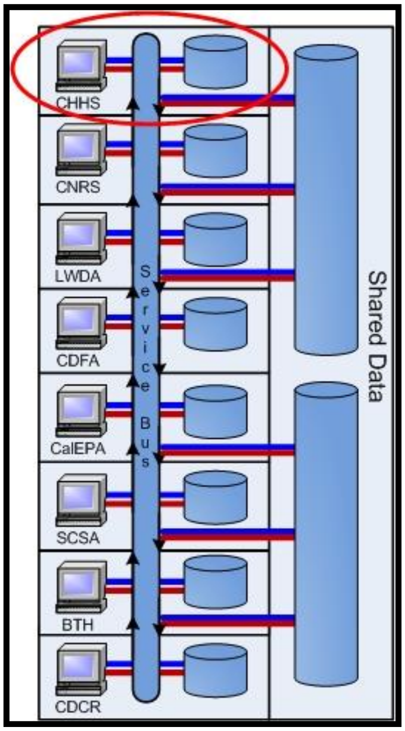


Figure 3-21 – Data Sharing Model

Figure 3-21 describes the model that is being recommended, where most data assets will be controlled by the agency. Data assets will be evaluated, and the ones that are valuable to the enterprise will be identified and will be a part of the shared environment. In this example, the CHHS agency uses their data assets as before and only those data assets that need to be shared with other agencies are in CDS.

3.4.9 Security and Privacy

Security defines the methods of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide integrity, confidentiality, and availability, whether in storage or in transit. Privacy addresses the acceptable collection, creation, use, disclosure, transmission, and storage of information, its accuracy, and the minimum necessary use of information. Primary drivers for security and privacy requirements are State/Federal/Local laws, organizational policies, regulations, market practices, contracts, and performance objectives. [xiv] The Data Strategy is aligned with the security principles identified in the following publications:

- Federal Information Processing Standards Publication (FIPS PUB) 199, *Standards for Security Categorization of Federal Information and Information Systems*,
- FIPS PUB 200, *Minimum Security Requirements for Federal Information and Information Systems,*
- Federal Enterprise Architecture Security and Privacy Profile *(FEA SPP),* and

- NIST Special Publication 800-53, *Recommended Security Controls for Federal Information Systems*

Currently, the agencies are independently working very hard to provide proper protection to their data assets. They are defining, incorporating, and maintaining necessary security and privacy policies. So when it comes to protecting "Shared Data", it would be imperative to carry forward and apply either the same or equivalent set of policies. To maintain a trusting and supportive relationship between all agencies that participate in the data strategy, a systematic approach towards a comprehensive security and privacy strategy needs to be put in place.

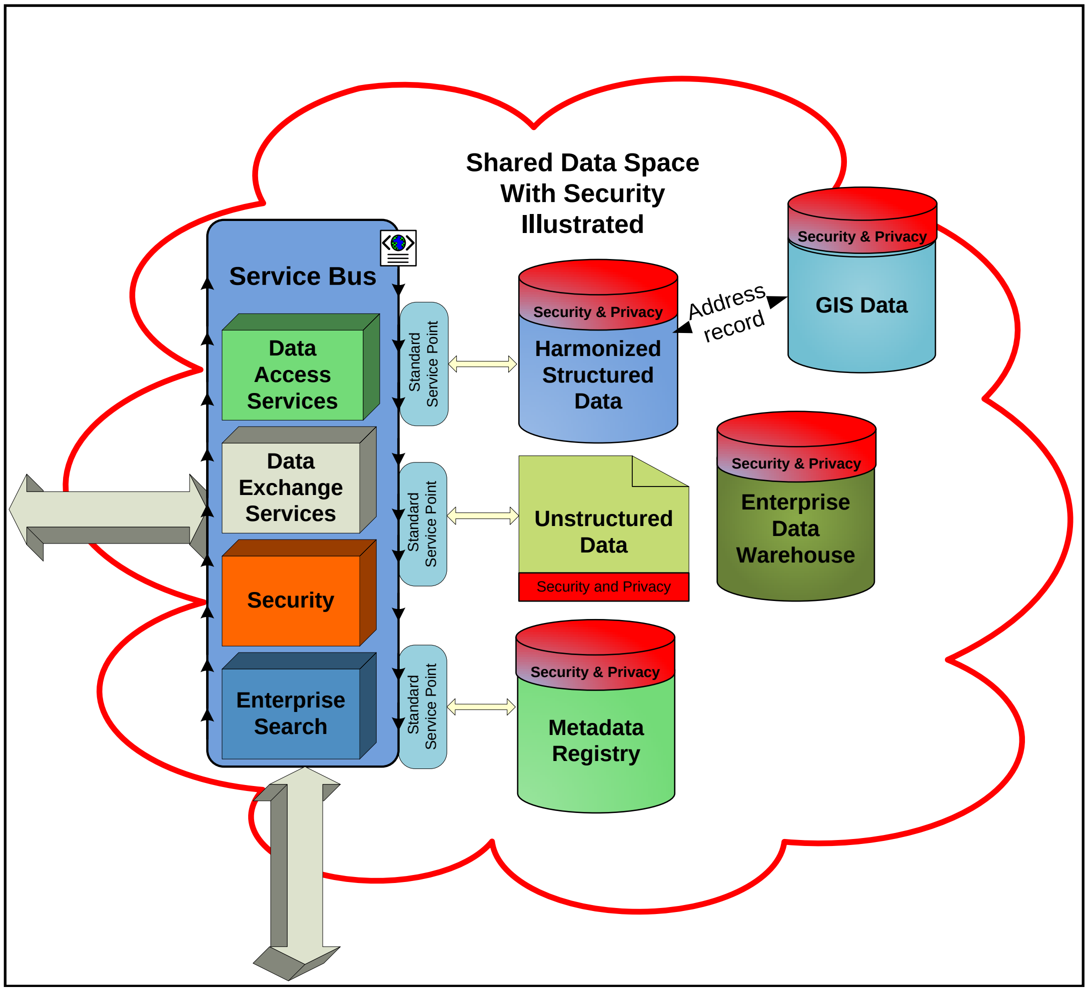


Figure 3-22 – Security in the Shared Data Space

The trust requirement could be met through a combination of agreements, advocacy, and technology. We will cover a systematic approach to designing a technical security architecture that encompasses security and privacy of "Shared Data".

3.5 Summary

The design and architecture of the Shared Data Space leverages a master data repository for structured data and enterprise content management application for unstructured data. The strategy for enterprise reporting and data warehousing is to provide a dimensional database within the Shared Data Space, and move more enterprise reports to the shared reporting infrastructure as more data is moved to the Shared Data Space. Standard interfaces to the shared data are provided using Service Oriented Integration and Web Services, and agencies can interact with these standard services to interact with the data.

The details around a California Trading Partner Network Services environment were discussed, and the importance of isolating external partner traffic was highlighted.

Finally, governance is the key for the success of the data strategy, and it will focus future direction, prioritize work, and resolve organizational roadblocks. Governance will also identify and coordinate data ownership, updated business rules, security requirements, and service level agreements for all types of data.

This page intentionally left blank

4. STRATEGY

This section provides an overall strategy towards achieving the implementation of “Architecture and Design” described in Section 3. It describes the project phases that need to be performed and the project methodology that is recommended with each phase.

4.1 Overview

The challenge with developing a data strategy for California is multifaceted. The State IT infrastructure landscape is extremely large and complex. Thousands of applications use widely varying technology from the past 40 years to support the State’s business processes. In our analysis, there are over 300 unique variations of software being used to support the State’s critical business processes.

This diversity of services, the idiosyncrasies of the legislation governing these services, and the varying state of the existing systems and data make it a challenge to decide where to begin. How can we get our heads around this complex set of requirements and work toward a data sharing strategy that is feasible? First, we must remember that implementation strategy cannot disrupt the delivery of these services. Then, we must discuss the assumptions around the strategy, evaluate how data is used across the State, and identify and coordinate the data sharing opportunities based on the risk and value to the State. The strategy is to start with the opportunities that benefit the State the most, while keeping the project phases small to reduce the risk and to ensure they are manageable.

Another consideration is ensuring that the shared environment is at least as secure as the agency environment. Without this assurance, agencies will most likely not buy into the shared data strategy. As security is such a key factor to the proposed architecture and its management, the key details are covered under Section 5.7.4.

4.2 Strategy Assumptions

In the early stages of strategy development, the desire is to keep risk at a minimum. Rather than proposing a few large project initiatives, the strategy is to provide an approach that can be delivered in many small manageable projects. According to the Standish Group 2004 CHAOS Report only 34% of all software projects are deemed on time and within budget. The report goes on to say “minimizing scope increases a project's chances of success. Minimized scope has replaced small milestones. While these two factors are similar, the act of minimizing scope leads to greater success than does creating small milestones”. We know from experience that the success rate for smaller IT projects is much higher than what is experienced for a single large IT project. Therefore, the recommendation is to implement the strategy in multiple small project phases. So what is the approach for segmenting the work into manageable projects? Where should we start, and how will the work be managed?

4.3 Approach

By leveraging the information provided to us by each of the agencies, we are able to understand the applications that are used by each critical service and the types of data used by each of the critical applications. By reviewing each application’s type of data, their description, and each data repository provided by the agencies, we were able to build an association to one or more Data Subject Areas. Each of these Data Subject Areas represents groupings of information, and each grouping breaks down even further into logical data models.

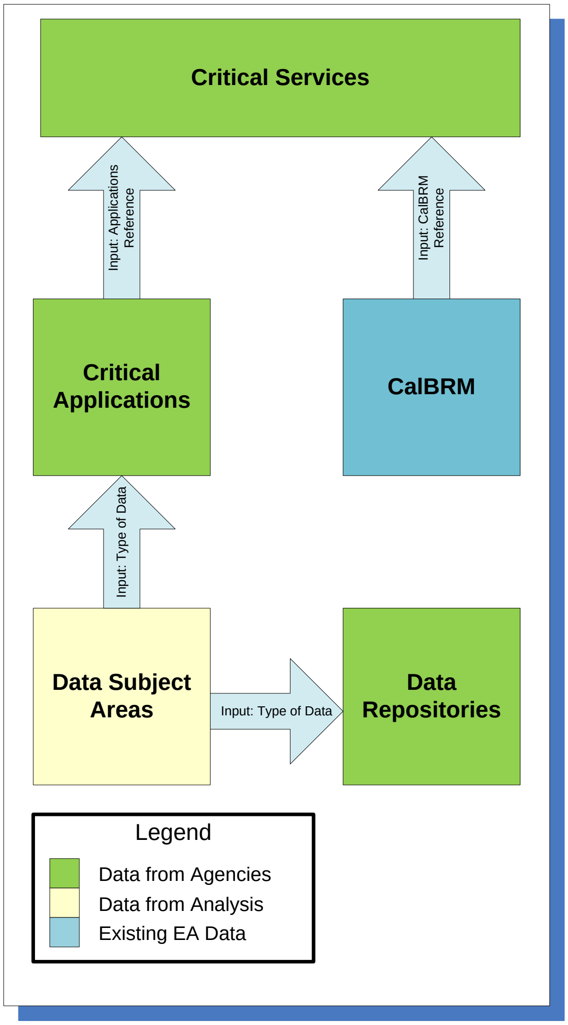


Figure 4-1 – Agency Data Structure

Understanding how the ‘business’ interacts with the data is just the first step in managing data and determining a data strategy. These interrelationships among the Data Subject Areas can be further explored. Figure 4-2 illustrates most of these relationships. Using this example, one can see that ***organizations,*** which employ ***people,*** use ***resources*** to produce a ***service***. These same ***services*** are provided to other ***people*** who may or may not be part of the ***organization***. Further analysis will be needed to further define the data subject areas and relationships that exist throughout the enterprise.

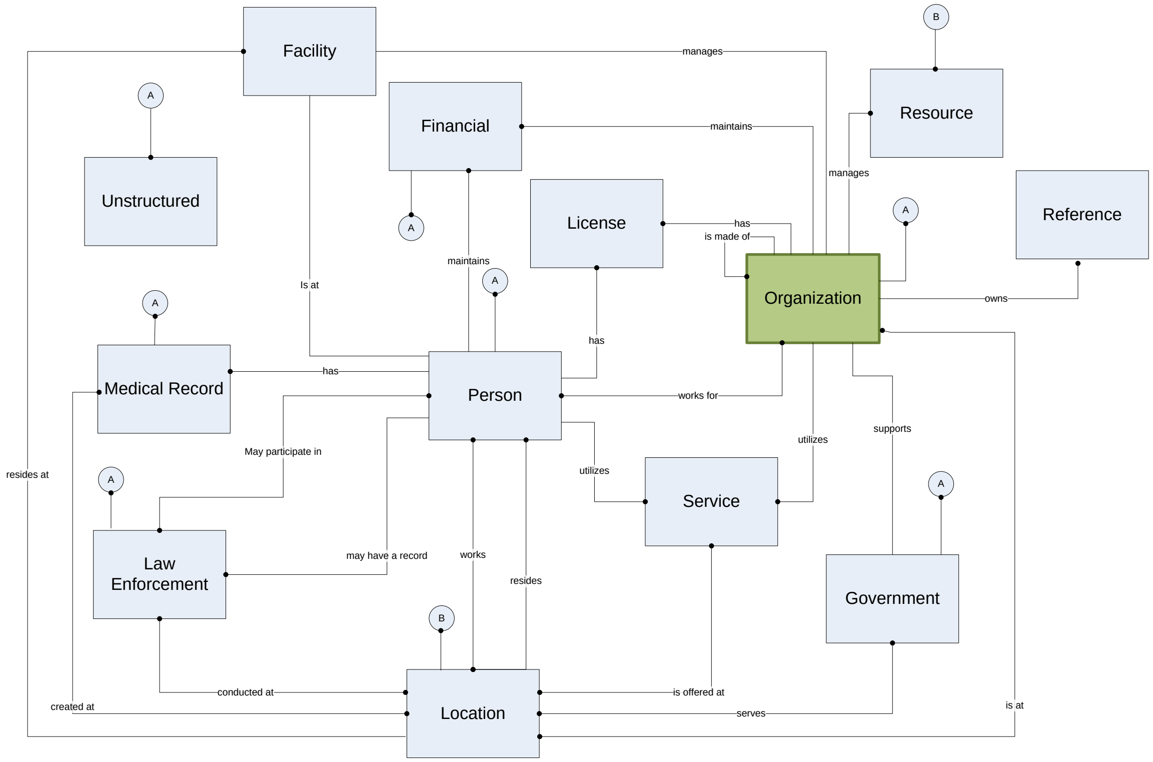


Figure 4-2 – Data Subject Areas

Each of these data subject areas (e.g., Organization) breaks down further into multiple entities, as illustrated by Figure 4-3.

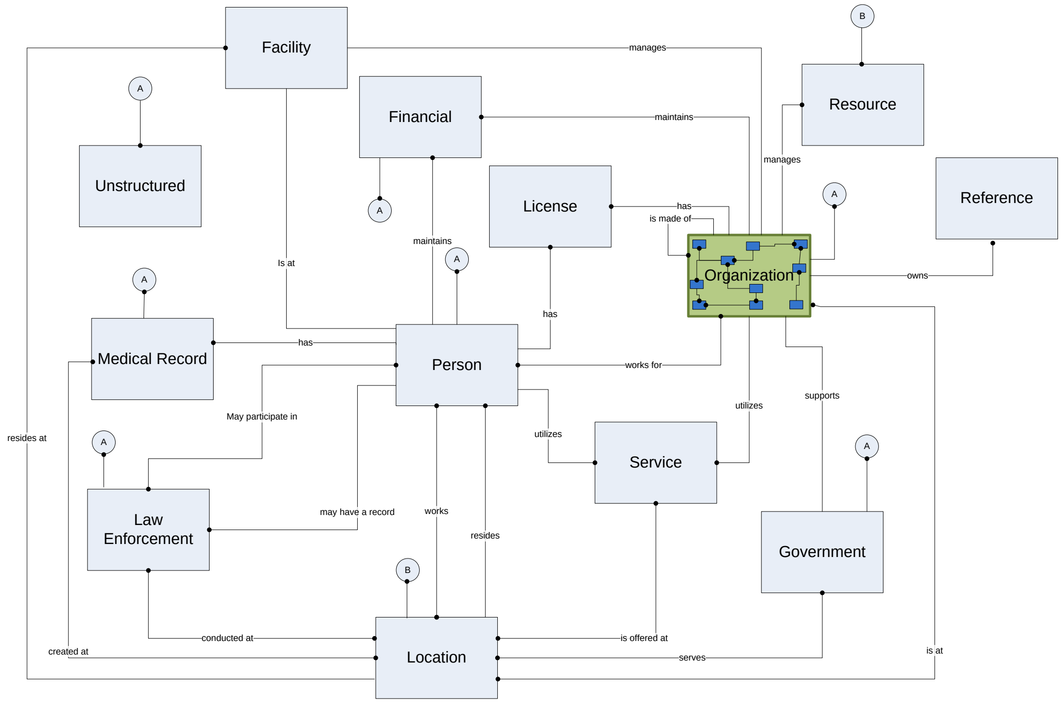


Figure 4-3 – Data Subject Area Decomposition

In this example, ***Organization*** breaks down into its own set of entities. Once these data subject areas were identified and associated with each of the critical applications, trends could be identified in the data. We were able to understand the information that each of the applications support, and got our first comprehensive view of the duplicated data within the enterprise. An opportunity does exist where third party data models can be leveraged for the State, aiding the decomposition of each subject area. *OASIS Customer Information Quality* is one example of a set of third party data models that may be leveraged. The fit analysis required to determine if these data models should be used is outside of the scope for this strategy.

Taking this process further, we linked the Business Services to references within the CalBRM (as illustrated in the Data Relationships Diagram). Each of these services has a relationship to applications that support a service. Leveraging this relationship, we can now see where the data touches the sub-functions within the CalBRM.

These relationships will allow us to see where the data impacts the business and will support future impact assessments.

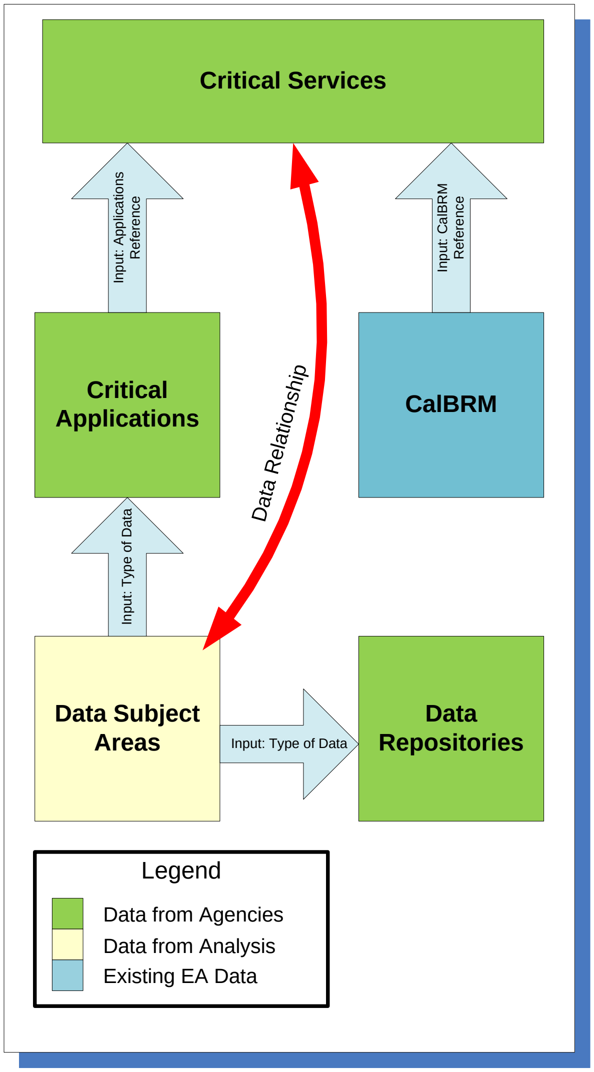


Figure 4-4 – Relationships within Agency Data

4.3.1 State Government Data Usage

During the analysis phase of the project, agencies reported on types of critical data and the applications that housed this data. Based on these reports, a good starting place can be determined by analyzing which data is most valuable to share. With the key data identified, the next step is to evaluate how the data is used for the applications, and the best possible business-use case that would engage a pilot group. Most of the applications were identified as high priority applications, so there is little distinction that can be made by the types of the applications. Table 4-1 illustrates rolled up counts of applications by supported data subject area in the agencies included in the survey.

Application Count	Data Subject Areas
358	Person
281	Organization
193	Financial
127	Government
127	Location
101	License
96	Unknown
58	Unstructured
40	Medical Records
38	Law Enforcement
37	Reference
36	Services
30	IT
27	Facility
21	Resource
7	Regulations

Table 4-1 – Agency Data Subject Areas

From this list we get an indication of the value each type of data has for the overall enterprise. From this list, the most prevalent subject area appears to be Person. ***Person*** would appear to provide the biggest benefit to a majority of the agencies and departments[17]. Although extremely common across agencies, this is not the recommended path as a starting strategy. The analysis and feedback revealed structural challenges that alone provide too many obstacles that should be avoided in the pilot implementation. The second most common *Data Subject Area* would be ***Organization***, while ***Financial*** information is the third. The fourth most common Data Subject Area is ***Location.*** The category of ***Location,*** however, may be a bit understated, as some agencies may group a field like address in more than one Data Subject Area. For example, address was assumed in some of the details relating to ***Person*** and ***Organization***. This is based on the table structure of the actual applications. As we did not analyze data and applications at a data base structure level, in our model we normalized it out. This is noted, however, as some agencies might have assumed it was a part of the Person entity.

Another area of focus is the ***Government*** subject area. Although ***Government*** subject area was more of a catchall, from the discussions from the agencies, it was determined that a solid understanding of a ***project*** location was of high interest to many of the

[17] Person may not be the best place to start due to outstanding challenges in how the data is organized. Probably the biggest challenge is identifying people consistently across California without using their Social Security Number. More of this issue is discussed in the Consideration subsection in this section of the Data Strategy Report.

agencies. For example BTH, CNRA, and CalEPA make heavy use of the concept of a ***project***, and knowing where these projects are across the State is of high interest.

Data Subject Area	Description
Facility	A building or a group of buildings
Location	A physical location
License	A license, permit, or registration
Organization	A business or professional group
Person	Californian demographic information
Financial	Financial data
Law Enforcement	Crime, Criminals, and Law Enforcement
Regulations	Governmental regulations – usually static data
Government	A bit of a catchall for anything that is used to run the government – includes projects, grants, testing, contracts, special IDs, etc.
Unstructured	Any data like word documents, spreadsheets, engineering drawing, etc.
Services	Any services offered to a Californian
Medical Records	Medical records
Unknown	Unknown
IT	IT Related
Resource	Physical Resource like a vehicle
Reference	Reference information

Table 4-2 – Data Subject Area Descriptions

4.3.2 Pulling it All Together

Successfully building a scalable statewide data and application-architecture requires a strong understanding of the business, the business process, and the data that is being used within the business process. To gain such an understanding will require documenting the following:

- Business processes
- The data used at each step of a process
- The interactions between both data and business processes
- The business drivers

Once we understand how the business interacts with the data, we should settle on and document the following:

- A standard data structure
- A standard means for interacting with the data (our recommendation is using web services)
- Industry standards to be used

- Data ownership
- Data lifecycle and update rules
- Data security
- Data auditing requirements
- A baseline directive for the service level agreements that will be needed to support the availability requirements

This information can be obtained in many different ways. Surveys and general feedback meetings are a nice start, but the most efficient means is face to face meetings with the business owners. Once the information is secured, this information is usually rolled into a modeling tool like Erwin Process and Data Modeler.

Once this information is gathered, it can be cataloged and submitted to the Data Governance Committee for review. The Data Governance Committee, covered in Section 5, will evaluate the documentation, and determine what should be part of the initial rollout for the data sharing solution.

4.3.2.1 One Size Does Not Fit All

Please note that the approach may vary depending on the type of data being stored. This is due to the uniqueness of the business of running California State Government. Finding an existing database structure that fits the defined business needs perfectly is unlikely. A custom database structure should be considered, as Master Data Database design is not the area to cut corners during your design phase as any shortcut is likely to cause more work in the long run.

On the flip side, however, finding an enterprise Document Management system that will not only meet but exceed the State’s requirements is a strong possibility. Therefore, the recommendation is to produce a custom Master Database for the master data records and to purchase an enterprise document management system.

4.3.2.2 Where to Begin

The State of California supports thousands of business processes. Their support drives the formation of a complex IT landscape. This landscape, coupled with hundreds of diverse technologies, provides a real challenge for where to begin a data sharing initiative. For the State of California, selecting the wrong starting strategy could be the beginning of the end. Failure is likely unless the strategic vision is tangible (i.e., agencies can easily relate to the business case). Failure could also result from too many agencies in the pilot or an overambitious initial project. Therefore, determining the starting point is critical.

A review of the trials and tribulations of other states’ statewide data sharing initiatives, led us to discuss a possible pilot partnership with California Geospatial Information Officer (GIO). One of the GIO’s first priorities will be to create common GIS data sets of imagery, roads, and landmarks of the State. It was suggested that Location be used as the data subject area to share first. This is one data subject area that touches all agencies, and has been a focus of target fixes with the GEO Spatial team. Eighty percent of the State's data has a spatial component, and although it is many times treated differently, there is no difference between GIS data and any other type of data.

Location can be described in many ways geospatially. One way is to describe it as a point in space (e.g., a latitude and longitude). Another way is by a physical address. The strategy is to support the California Geospatial Framework where a location can be described by as many as 18 different ways. The strategy supports traversing these geospatial layers in any direction. For example, with a geospatial point, the corresponding mailing address can be found, and with the mailing address, the point related to any of the 17 other layers can be found. Both conditions have enormous value to the state. In addition, having a set of valid locations is a huge asset for any agency. Some examples where validated locations are important:

- The Franchise Tax Board receives tax returns and mails refunds to locations throughout the State. A valid mailing address is critical.
- The California Environmental Protection Agency tracks facilities and waste with respect to population areas.
- The California Natural Resource Agency tracks resource assets across the State.
- The California Health and Human Services distributes welfare, licenses nursing home facilities, and tracks services offered to California's constituents.
- The California Department of Food and Agriculture tracks pests, dairy farms, and farmers.
- The State Consumer Affairs Agency licenses businesses and provides permits.
- California Department of Corrections tracks facilities, inmates, families, and parolees.
- The Employee Development Department maintains the location of job seekers and employers.
- The Board of Equalization processes twenty million transactions per hour relating to sales tax. The location of the purchase is important and errors occur five percent of the time while processing the tax.

The examples are endless. Since most examples relate to something physical, a validated location becomes important. In addition, 'Location' data subject area supports most of the interrelationships that are tracked across agencies. A few of examples are:

- CHHS may want to know where nursing homes or hospitals are located with respect to an EPA toxic facility.
- Fraud can be tracked within an area by looking at data trends related to per capita averages throughout the State. For example, a spike in welfare recipients may be seen in data for an area that exceeds the State averages, leading to further investigation.
- SCSA may want to track earthquake fault lines as they relate to new building permits as they relate to the type of business.

Many of these questions can be asked today, but data must be gathered from the agencies and the complex interrelationships must be built.

Now that the base line starting point is identified, ***location***, the following additional steps will be applied:

- Define Business problem
- Define Business solution
- Define Business use and impact
- Identity at least two sponsoring agencies
- Funding *(Based on the current economic conditions, the initial funding will need to be established for the pilot teams. Without funding at the start, the likelihood of moving forward is not favorable).*
- California Department of Public Health (CDPH) already has geocoding in its strategic plan. It is desirable to leverage their analysis and provide some alignment with their initiative.

4.3.2.3 Address

As mentioned earlier, knowing the specific location of people, businesses, facilities, projects, and services offered is very beneficial to the State. There are several ways to describe a location. One way is by a point (e.g., x, y) using a coordinate system and another is by a physical address. There are two types of addresses that are being maintained with the data. They are:

- **Situs** address (Latin for position or site)
- Mailing address

The Situs address is the address that has been assigned to the parcel, and is a part of the GIS information. The Situs address may not be the mailing address. However, the Situs address is useful for Emergency Response and other location identification. Since there are times it is not the mailing address, it cannot be relied upon for communication purposes.

The mailing address is what is used by most agencies, and is used to contact businesses and people. The recommendation is to maintain a listing of valid mailing addresses, but to tie them back to the GIS information that exists for the parcel.

Since the GIS project is a different initiative, the strategy calls for supporting the necessary information about a mailing address so that the corresponding GIS information can be determined. To do this, we leverage the Federal Geographic Data Committee - Street Address Data Standard[xv]. Right now it is a version 2.0 draft. The attributes that actually provide this linkage are as follows:

- Address X Coordinate
- Address Y coordinate

- US National Grid Coordinate
- Address Latitude
- Address Longitude

For more information on each of these elements, refer to the Federal Geographic Data Committee - Street Address Data Standard. A copy of the standard is attached in Appendix H. The US National Grid Coordinate (USNG) provides the context for the Address X and Y Coordinates.

> "US National Grid (USNG). This standard established a nationally consistent grid reference system, just as all street maps use a common set of street names. USNG provides a seamless plane coordinate system across jurisdictional boundaries and map scales; it enables precise position referencing with GPS, web map portals, and hardcopy maps. Unlike latitude and longitude, the USNG is simple enough that it can be taught and effectively used at the 5th grade level. It enables a practical system of geoaddresses and the universal map index."
>
> FGDC USNG Information sheet 4. http://www.fgdc.gov/usng/USNGInfoSheetsCv5_4pages.pdf

The USNG is defined in document FGDC-STD-011-2001[xvi].

Latitude and longitude can be used as an alternative means of addressing a point in space is, and these are also called for within the Street Address Data Standard. All other address fields are called out in this data standard.

4.4 Project Methodology

Since the scope of the work is so large, it is recommended that the work be broken up into multiple project phases. Some of these phases should be executed using a waterfall project management methodology while others should use an iterative approach, such as a spiral project methodology. An example of this would be the initial build out of the shared infrastructure. An iterative approach for the initial build simply does not make sense. However, an iterative approach to building web services not only makes sense, it also is recommended. Another factor is timing of the phases. For example, the infrastructure must be in place before much of the master repository work occurs.

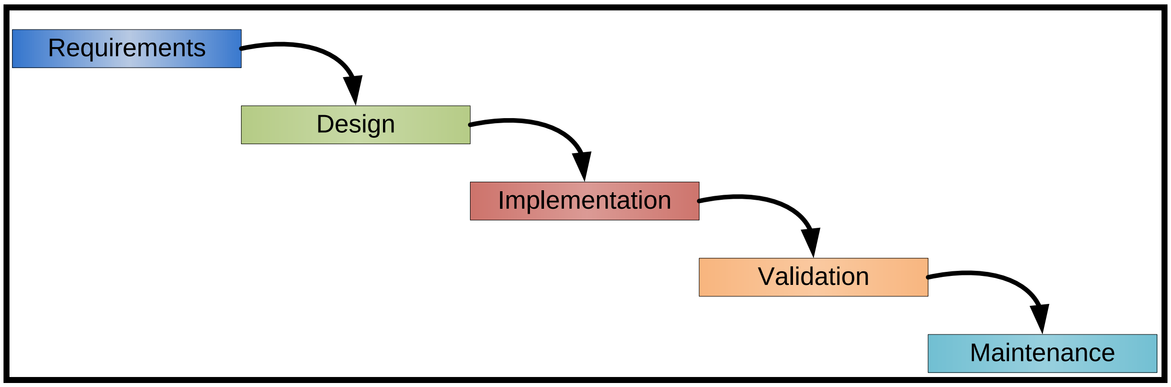


Figure 4-5 - Waterfall Methodology

For phases that are executed linearly, a traditional waterfall project methodology is recommended. This methodology has been used for many years and has an execution plan that builds on the previous project phase. Requirements are gathered, a solution is designed, implemented and tested. Finally the solution is made available to the business users and goes into a maintenance phase. The benefits of a waterfall methodology are; it is well understood by project managers and provides excellent project control and documentation. A few of the limitations with a waterfall project management methodology are; it does not lend itself to iterative development processes and the project scope must be well understood at the onset.

We recommended that phases that can be performed iteratively be organized into multiple smaller phases called spirals[xvii]. The scope of each spiral will be limited allow a functional and useful result to be produced and released in six months. Each spiral can be managed separately with some coordination to the others. Each spiral will implement new functionality and/or enhance functionality from a previous spiral. A duration of six months was chosen as an appropriate timetable for the State of California. Shorter spirals limit the scope too much and do not allow the State agencies enough time to participate in a project of this sort. Longer spirals will define a too large of a scope. Like with any endeavor, progress is the key and showing usable functionality early is vital.

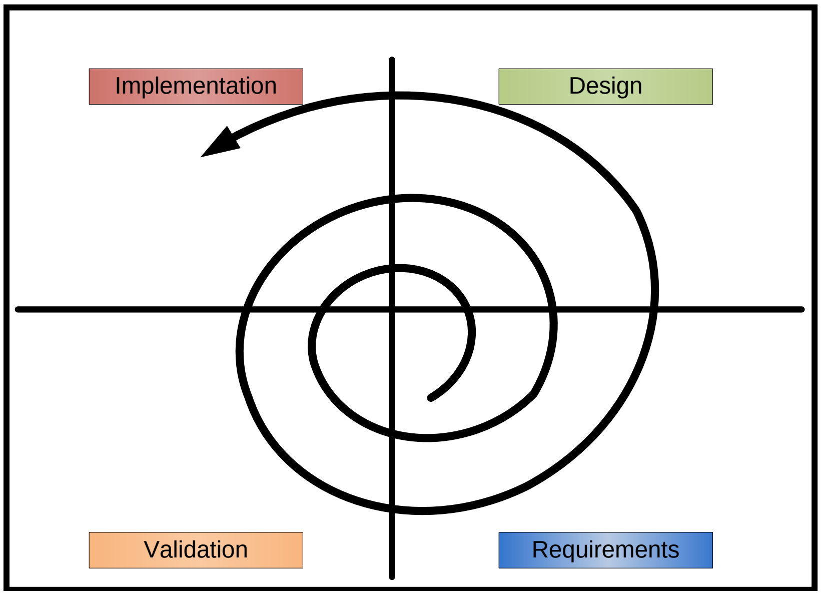


Figure 4-6 - Spiral Project Management Methodology

This spiral approach allows the shared space to be incrementally expanded while also allowing any deficiencies in design or functionality to be addressed in the next spiral. This is an ongoing project with a multi-year commitment; however the benefit of using spiral development process is that the State will be able to use the built functionality after about 1 year. Supporting this flexibility has a liability. Once the shared space is in use, future extensions and updates to the shared space may impact State business that is currently using the Shared Data Space. Therefore, a robust testing cycle needs to be performed to ensure a quality release from the onset, and rigorous change management of the environment must be in place.

4.4.1 Architectural Components

CDS will be addressed in nine project phases. From a strategy perspective they are interdependent with one another. The security architecture will define the security standards and overall security design that will be leveraged by the other phases. The Enterprise Infrastructure provides the base hardware and software that is used by most of the other phases.

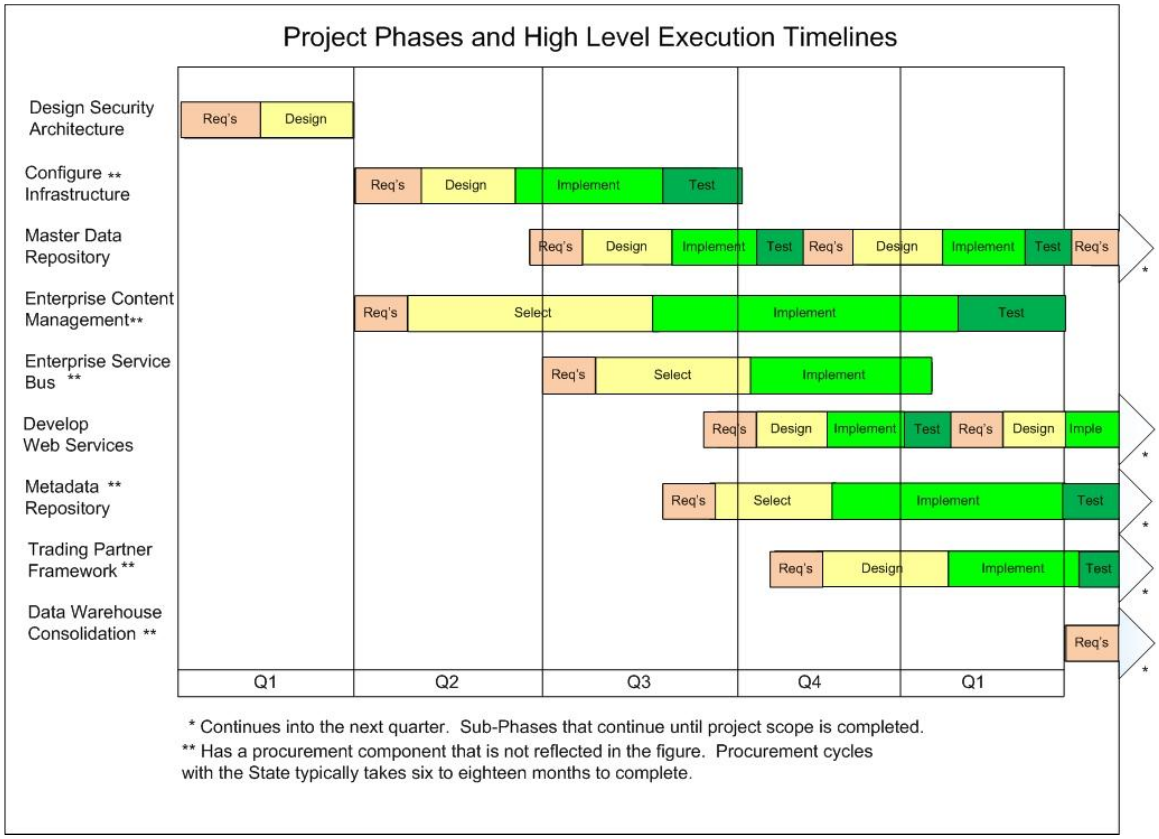
The project phases are:

1. Security Architecture Design
2. Enterprise Infrastructure
3. Metadata Registry
4. Master Data Repository
5. Enterprise Service Bus
6. Web Services to Interact with the Master Data Repository
7. Enterprise Content Management System
8. Trading Partner Network

9. Data Warehouse Consolidation

A brief discussion is given on the specific approach to each of these components, as their implementation approach is very different. In these sections, project methodology, whether a waterfall or a spiral approach should be taken, and a high level timeline and dependencies are identified. Planning for security is addressed in its own project phase however the implementation of the security plan is addressed individually within the project phase for each of the components.

The timelines in Figure 4-7 are for illustration purposes, but represent the projects and durations needed to establish the data sharing environment. The procurement process is not reflected in the timeline in Figure 4-7 as it can stretch out several years. These project phases are described in more detail in the following sections.

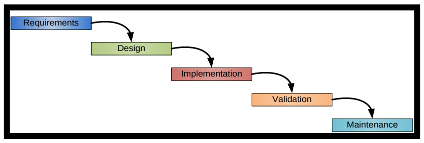


* Continues into the next quarter. Sub-Phases that continue until project scope is completed.
** Has a procurement component that is not reflected in the figure. Procurement cycles with the State typically takes six to eighteen months to complete.

Figure 4-7 – High Level Timeline

4.4.2 Data Strategy Overall Project Approval and Procurement

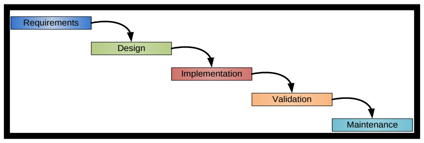
Project Methodology:	Waterfall
Timeline:	Six to 18 months
Start:	Immediately



The work required for the initial concept, Feasibility Study Report (FSR) and the accompanying Budget Change Proposal (BCP) must be completed first. Due to the comprehensive procurement process that exists with the State planning for the initial project procurement must be performed upfront.

4.4.3 Design Security Architecture

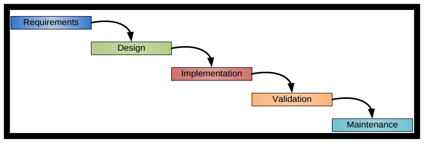
Project Methodology:	Waterfall
Timeline:	Three months
Start:	Immediately



Security for all sub-phases for CDS must be designed from the ground up. The overall security approach will need to be considered in the light of each of the project phases and the technology that was selected. Once the security requirements are identified, design completed, and test cases written to validate the design, this information will be passed along to each of the project phases to ensure compliance.

4.4.4 Configure Infrastructure

Project Methodology:	Waterfall
Timeline:	Six months
Start:	After Security Architecture Design



The infrastructure for the system includes application servers to support each of the components; the cluster database servers to support the master data repository, the enterprise content management system, the metadata repository, and the enterprise data warehousing solution. In addition, all of the software licenses to support the system must be procured.

The infrastructure must be able to support operations 24x7, however, at the onset a longer maintenance window can be supported. The recommendation is to start small and evolve the solution to support more and more functionality. This small start allows some room for flexibility in the hardware design as the data needs to evolve and progress over time. In addition, a minimal configuration can be instantiated to support the initial business, and then grown.

The initial minimal configuration is seen as encompassing the following:

- Redundant Load Balancers
- Redundant Web Servers
- Clustered Application Server
- Clustered Database Servers
- Highly Available Tiered Storage
- A Server Management Console

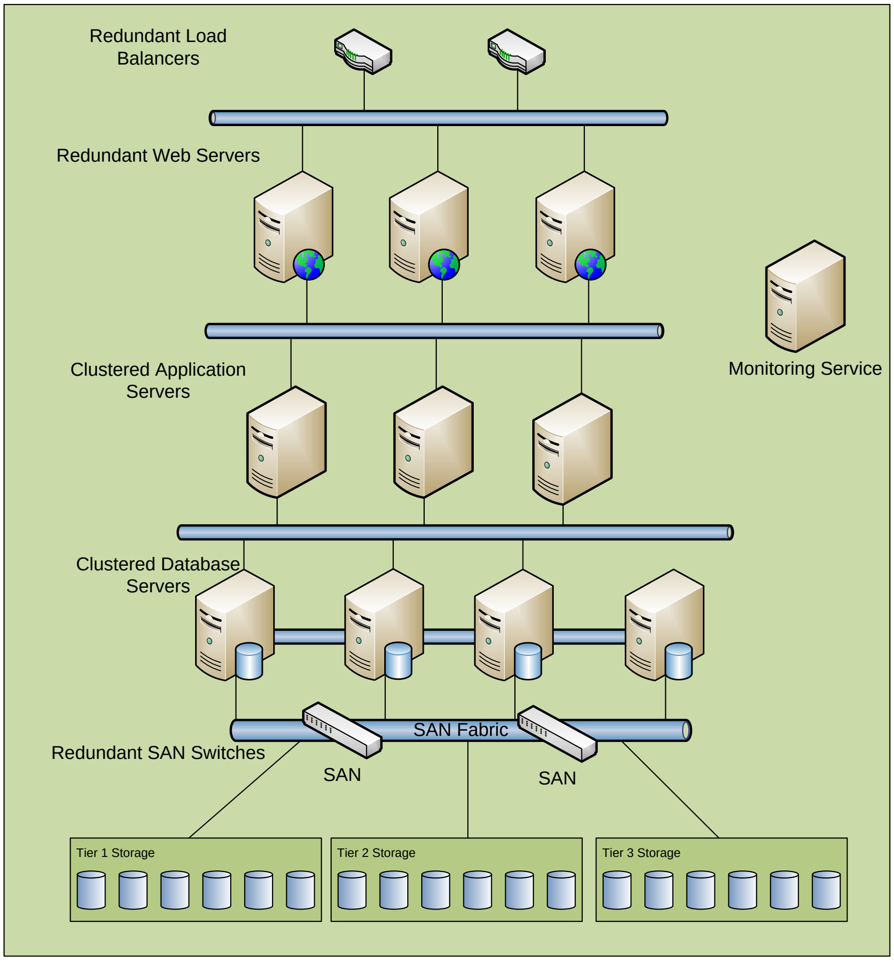


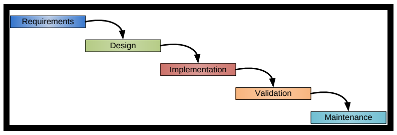
Figure 4-8 – Minimum Configuration

To estimate the cost and the work required some assumptions have been made for the initial environment. Figure 4-8 is for illustrative purposes, the number of servers in each tier will need to be evaluated to ensure the environment is sized properly.

Preparing for a disaster is always important. However, the directive for the initial pilot implementation is that a disaster recovery (DR) site can be added to the solution at a later date. When more departments leverage this solution or if a department requires higher availability then a disaster recovery site will need to be in place. Depending on the disaster recovery solution that is implemented, the work and overall cost involved with the Infrastructure Configuration can more than double. Therefore it is recommended that the State implement the DR site only when it is needed.

4.4.5 The Metadata Registry

Project Methodology:	Waterfall
Timeline:	Six month Iterations
Start:	Immediately



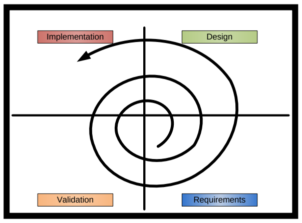
The metadata registry will need to be established. Though there are several metadata registries that are ISO/IEC 11179 compliant, there are very few vendors that currently offer solutions. Our research revealed only two ISO/IEC 11179 standard compliant solutions, they are.

- OneData Metadata Registry from Data Foundations
- Enterprise Metadata Manager from Oracle

We recommend evaluating the two solutions. Once the evaluation is conducted the metadata registry is to be established in the **shared space**. Any comparison of products may add time to the schedule.

4.4.6 The Master Data Repository

Project Methodology:	Spiral
Timeline:	Six month Iterations
Start:	Immediately

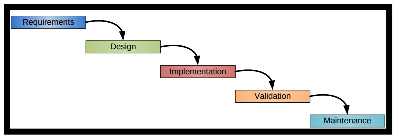


The master data repository will be taken in phases. The recommendation is to start small and build the repository, making sure the information is normalized so there is only one single point of truth. Consideration on the 'key' attributes for the data should be given to ensure multiple records do not represent the same physical entity. All duplications will eventually need to be resolved, and the best approach is to avoid this type of duplication to begin with.

Specifically, for the initial master data repository, address should be instantiated in the master data. All address information should be processed through a zip+4 address formatting engine to ensure that the information is valid. As mentioned earlier, address has significance when it is used in conjunction with some other data entity such as a person, business, facility, or service offered. Once the address is instantiated in the master data repository, the next phase of the project will be to identify and instantiate the related data entities.

4.4.7 The Enterprise Service Bus

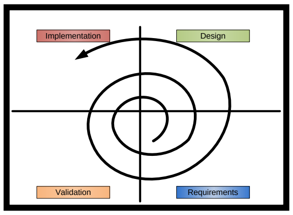
Project Methodology:	Waterfall
Timeline:	Six months
Start:	Immediately



Evaluating and implementing the enterprise service bus (ESB) within CDS will be organized into its own six month spiral. The goal is to choose an enterprise solution for the ***shared space*****,** not to dictate an enterprise solution for all agencies. An agency can then decide to use another standards compliant ESB solution with little or no functionality loss.

4.4.8 Web Services to Interact with the Master Data Repository

Project Methodology:	Spiral
Timeline:	Six month Iterations
Start:	As soon as the service bus is selected



There are two halves to a functional web service. They are called service points. In any service oriented project, there are always multiple service points that exist. In Figure 4-9, there is a ***standard service point*** that exists for interacting with the master data. The other service points (i.e., numbered service point 1 thru service point 3) interface the agency systems and the enterprise service bus. Since these systems are technically disparate, they must formulate the message that the standard service point requires. The strategy for the ***standard service point*** is provided for within this strategy, but the agency service points are not[18]. Since the agency applications have not yet been selected to interact with CDS, the strategy for these service points will need to be developed on a 'case by case' basis.

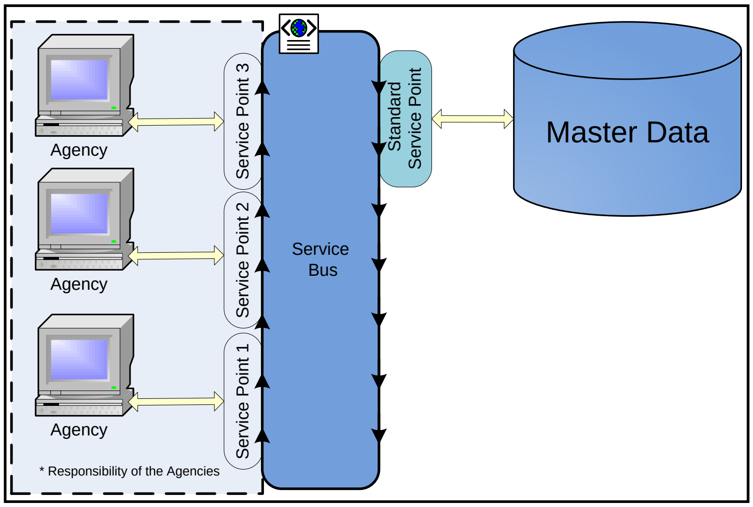


Figure 4-9 – Service Points

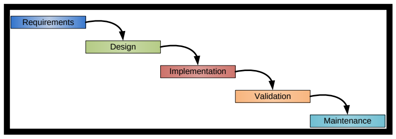
Work should also be accomplished in a spiral fashion after both the master data repository and the enterprise service bus are in place. The spirals for this portion of the work can be shorter than six months, and more of an Agile approach can be taken. Proper documentation should be created for each web service, since only the standard interface portion will be done within this project. To use the standard interface, the other service point, which is particular to an agency and its technology, will need to be built.

[18] It should be noted that a single service point is not functional. Data must be produced and must be consumed so it is fair to consider the minimum configuration for a service is to create both end points to the service.

Web Service	Description	Priority
setAddress	Creates or updates an address record.	1
getPointFromAddress	Pass a standard address, datum ellipse and projection model into it and it will validate the input and return an x value, y value. Datum and projection are optional.	2
getAddressFromPoint	Send it x value, y value with a datum ellipse and a projection and it retrieves the nearest address information associated with the address layer.	3
isAddressValid	It validates address and reformats it into a standard format.	4
isValidPoint	Validates x and y values against a datum and a projection model. Is it a valid point in California?	5
getGISLayer	When supplied a valid point it will retrieve one, multiple or all of the GIS framework layers that contain the point.	6
getDistance	Get the distance between two points, a point and a line or a point and a polygon.	7

4.4.9 The Enterprise Content Management System

Project Methodology:	Waterfall
Timeline:	Six months
Start:	Immediately



Enterprise Content Management systems have been evolving over the past few years, and are now considered mature products that are feature rich. Currently, the State of California has at least three different types of document management systems deployed at different agencies.

Content Management Tool	Agency
Microsoft Sharepoint	CalEPA, CDCR, CDFA, LWDA, SCSA
IBM FileNet	CHHS, LWDA
OpenText LiveLink	CHHS
EMC Documentum	CDCR
Oracle Stellant	SCSA
Ektron	SCSA
Custom Solution	CNRA
iManage	CHHS
Clarity	CHHS
InfoImage	LWDA
Pegasus	CDCR

Table 4-3 – Agency Ownership of Content Management Tools

As these systems are already in use, the recommendation would be to analyze these deployed systems to see if any of them meet the criteria needed for the proposed architecture identified in the statewide data sharing strategy.

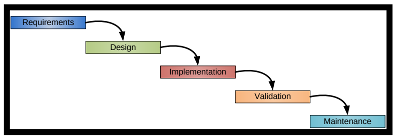
The products that merit investigation include:

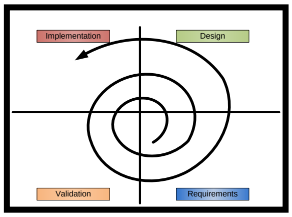
- Oracle Universal Records Management (URM) and Universal Content Management (UCM). It has a policy engine, metadata repository, and search capabilities. It may fit the need, and it is U.S. Department of Defense 5015.2-certified. The UCM component is a full featured enterprise content management repository.
- IBM Records Manager and Enterprise Content Management (ECM). The records management component has a policy engine and supports federated records management. It too is U.S. Department of Defense 5015.2-certified. The ECM component is a full featured enterprise content management repository.
- OpenText ECM Suite. Full featured Enterprise Content Management system. Its record management component is also U.S. Department of Defense 5015.2-certified.

All support a repository for semi-structured and unstructured data, and some organize and index data in third party content management tools. This is called federated records management.

4.4.10 Trading Partner Framework

Project Methodology:	Waterfall, Spiral Combination
Timeline:	Six months Increments
Start:	When needed



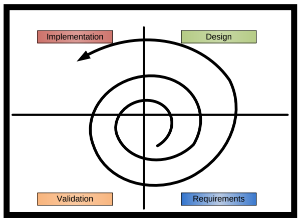


The trading partner framework consists architecturally of one or more application servers and a set of web services. It is recommended that the state build this once the need arrives. Currently, the agencies have a mechanism for interacting with entities outside of State government. The infrastructure component of the trading partner framework will be executed in a waterfall fashion or alternatively at the beginning of the first spiral. The externally facing web services should be created in a spiral fashion.

This initiative should only be started once the need arises and once the Shared Data Space can supply or receive the data for the external interface.

4.4.11 Consolidated Data Warehouse

Project Methodology:	Spiral
Timeline:	Six month Iterations
Start:	After a few iterations of the master data has been performed.



The consolidated data warehouse initiative should also be performed in spirals. The subject of the work that is to be performed within each spiral will be closely aligned with the Master Data initiative but delayed by one or two spirals. An enterprise data warehouse can be built using data directly from the agencies, however, the real time and effort savings comes from using data within the Master Data Repository.

Currently, some agencies have their data warehousing solutions in place. As data is migrated into the Master Data repository to be shared, it then becomes a candidate for the data warehouse. Standard ETL jobs can be built to extract and transform the data from a normalized format in the Master Data Repository into the dimensional structure needed to support reporting requirements. One-off ETL jobs can be used to extract data from an agency as needed to complete a reporting requirement. These one-off ETL jobs should be minimized.

4.5 Next Steps Considerations

The following subject areas should be addressed in the following order:

1. Organization
2. Facility
3. Services
4. Public record data
5. Personal data

The final goal is to have an accurate and comprehensive view of each of these areas, including person. To understand how services are offered, to whom, and where across the State is vital for accurate decision making.

There are a few key factors for consideration when sharing data about a person. Among the agencies reviewed, the detailed data analysis revealed that there is no common unique identifier for person. At surface value, the most obvious choice is social security number. A social number cannot be used as a unique identifier for person for two key reasons. First, there is the potential issue of identity theft. The second is that a Californian receiving services might not be a U.S. citizen, and thus lack a social security number.

Another consideration is the privacy statutes and laws that are in place. Privacy is a big issue for any government entity, especially when it relates to a person's privacy. California State Statutes 1798 details the obligation that is on the state to maintain a person's privacy. The highlights of this statute include the following topics:

- **Openness** – Disclosure about the existence of a data store containing personal information and main purposes of use. California Statute 1798.16a
- **Collection Limitation** – Data must be obtained lawfully and with prior consent or knowledge of the person who is the data subject. California Statute 1798.17
- **Use Limitation –** Personal data cannot be used for any other purpose than those specified in the ***collection limitation.*** California Statute 1798.17
- **Data Quality –** Personal data should be relevant to the purpose it is used. California Statute 1798.18
- **Individual Participation** – A person has the right to obtain permission from a data controller the ability to update their data. California Statute 1798.16c, 19
- **Security Safeguards** – Personal data must be protected. California Statute 1798.21
- **Accountability –** A data controller must be held responsible for the above practices. California Statute 1798.19, 20, 21, 22

For a more exhaustive discussion on this statute and other privacy requirements please refer to the Office of Information Security and Privacy Protection (OISPP). Their website is http://www.oispp.ca.gov. There are several ways to meet the statues and still make personal information available to all agencies in the State. They are:

- Inform people of the existence and function of the Shared Data Cloud and on every form they provide personal information make sure they know it will be shared with all other agencies.
- Pass a law similar to Colorado Law ***HB 08-1364*** *that revises the State statute, Act* Article 37.5 of title 24. This law gives the agencies the ability to share a certain data elements about a person.
- Provide a website that allows constituents to update their details that are known within the State. Use of the website identifies consent to use the information within all applicable agencies.

Even though this issue is addressable, maintaining information about a person can still be controversial. The idea of the data strategy is simply to improve the quality of the data that is being used in the State and hence improve decision making by state leadership.

4.6 Summary

The proposed approach is to define a small manageable scope around commonly used data that adds the most value to the state. The identified data set is address. Once valid address information is made available to the agencies, other related data sets (e.g., Person, Business, Facility, or Service) will be identified and added to the Shared Data Space as per the direction of Data Governance Committee. Each initiative will be addressed as a small project lasting roughly six months. This approach ensures steady progress while minimizing overall project risk.

This page intentionally left blank

5. ORGANIZATIONAL CHANGES

This section offers recommendations towards organizational changes in achieving statewide data sharing, consolidation of data, and overall governance related to the strategy. Several committees are recommended and the process for managing change to the environment, to the business, and to govern the data is discussed. In addition, departmental changes that are required to support the data strategy are identified.

5.1 Overview

A critical aspect to any proposal is managing the change impact to the business, and IT is no exception. As technology is introduced, new choices are available to the 'business' that were not originally available. These choices can have both positive and negative consequences. All change must be thoroughly evaluated, understood, and managed to promote the positive and avoid the negative. The State should establish a Change Control Committee for the sole purpose of managing the change required for sharing data within the state.

Governance is about understanding the impact decisions will have on the 'business' well before the decisions are implemented. Since the State of California's business is complex, the business impacts of decisions that are made are equally complex. For this reason, good governance requires good representation of the business in a decision making process based on solid business processes. The State should establish a Data Governance Committee to manage the future direction of data sharing within the State of California.

Finally, organizational changes will be needed within OTech to support the shared infrastructure. Several concepts have been brought out and discussed with respect to mainstream adoption by the IT datacenter.

5.2 Change Management

Any major undertaking changes how business is performed and supported. With the creation of such a critical business asset, the central data repository, changes to the organization will be necessary to support the solution. To achieve statewide data sharing, the OCIO will have to embrace organizational change management. Change management is a structured approach to transitioning individuals, teams, and organizations from a current state to a desired future state.

Change management includes processes and tools for managing the people side of the change at an organizational level. These tools include a structured approach that can be used to effectively transition groups, teams, or agencies through change. When combined with an understanding of individual change management, these tools provide a framework for managing the people side of change.

Organizational change management processes include techniques for creating a change management strategy (readiness assessments), engaging senior managers as change leaders (sponsorship), building awareness of the need for change (communications), developing skills and knowledge to support the change (education and training), helping employees move through the transition (coaching by managers and supervisors), and methods to sustain the change (measurement systems, rewards, and reinforcement).

5.2.1 Management's Role

Management's responsibility is to detect trends in the overall environment, enabling their team to be able to identify changes and initiate programs. It is also important to estimate what impact a change will likely have on employee behavior patterns, work processes, technological requirements, and motivation. Management must assess what employee reactions will be, and craft a change that will provide support as workers go through the process of accepting change. The nurturing and support of the change environment will then promote the implementation of said changes.

5.3 Change Management Principles

Adopting a principled approach that displays integrity and engenders openness and trust will see your change program through the hard times. Some common key principles of successful change management are suggested below. These principles are:

CHANGE MANGEMENT PRICIPLES	
Sponsorship	The change program has the visible support of key decision-makers throughout the organization, and resources are committed to the program
Planning	Planning is conducted methodically before program implementation and committed to writing. Plans are agreed with major stakeholders, and objectives, resources, roles, and risks are clarified.
Measurement	Program objectives are stated in measurable terms, and program progress is monitored and communicated to major stakeholders
Engagement	Stakeholders are engaged in genuine two-way dialogue in an atmosphere of openness, mutual respect, and trust.
Support Structures	Program implementers and change recipients are given the resources and supporting systems they require during and after change implementation.

Table 5-1 – Change Management Principles

5.4 Change Program Stakeholders

Success of the OCIO's change program will depend upon a range of people. These people can be divided into five stakeholder groups. A stakeholder is any person with an interest in the process or the outcome of the proposed organizational change. Consider each group separately if you are to avoid one or more groups falling off the edge of the map just when you find that you need them the most.

Table 5-2 – *Stakeholders* provides a description and examples of each change program stakeholder group.

Stakeholder Group	**Description**	**Examples**
Change Recipients	The intended receivers of the products of change or change outcomes.	Pilot agencies Agency departments Public web access
Decision Makers	The people that approve a change effort and decide its scope and direction. In the proposed governance model, they are considered the combination of the Executive and strategic levels of the governance model.	Steering Committee Members Project Sponsor Chief Executive Officer
Resource Holders	The people authorized to release financial and human resources required by a change effort. In the proposed governance model, they are considered the legislative model.	Agencies OCIO OTech
Executers	The people charged with the responsibility for bringing about the change. They are the program implementers.	Program Manager Project Manager Project Team Members
External Parties	The people that are not the intended recipients, but who are impacted by the change.	WEB portal users whose access to a business is restricted after a change in business hours

Table 5-2 – Stakeholders

Some of these identified stakeholders listed in the table above are crucial to the proposed governance models.

5.4.1 Communication

Once you have identified your stakeholders, consider the key messages you will need to deliver to each group in order to gain their support. You will need to tailor your message for each group, showing them the value added by the change.

Once you have identified your key messages for each stakeholder group, you will need to identify the best way to communicate these messages. Consider the communication style and preferences of each target group. Possible communications styles and types to be used are: Email, phone calls, update meetings, 1:1 face-to-face, and web page updates. Some of these methods and modes of communication will suit some stakeholder groups and are not acceptable for other stakeholder groups. Seek to find the best approach and tailor it appropriately.

5.4.2 Change Management Approach

A structured approach for change management can be helpful. One such tool was developed by Business Performance Pty Ltd. This phased approach can be applied to any organization and is easy to remember. It is referred to as the **CHANGE Approach**. This approach consists of six phases that successful change programs progress through.

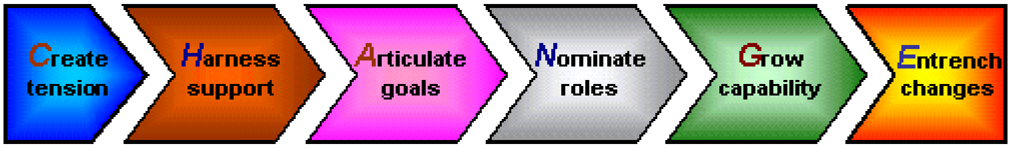


Figure 5-1 – Change Process Flow

Create tension	Articulate why change needs to happen and why it needs to happen within the planned timeframe.
Harness support	Get on board the key decision-makers, resource holders, and those impacted by the change.
Articulate goals	Define in specific and measurable terms the desired organizational outcomes.
Nominate roles	Assign responsibility to specific individuals for the various tasks and outcomes.
Grow capability	Build organizational systems and people competencies necessary for effecting the change.
Entrench changes	Institutionalize the change to make it “the way we do things around here”.

Table 5-3 – CHANGE Process Steps

5.4.3 Change Management Execution

Business change management should be managed by the business itself; however, there may be architectural and technical aspects to the changes that need to be managed as well. Therefore, it is recommended that the State adopt the concept of a Change Management Committee with representation from the following areas:

a. Business executive from every agency

The role of business executive is to provide input towards current business processes and the restrictions around those processes such as laws, policies, regulations, contracts, and performance objectives.

b. Enterprise Architect from every agency

The Enterprise Architects would be responsible for decomposing business processes and recommending solutions.

c. Information Security Officer from every agency

The Information Security Officer (ISO) is responsible for evaluating and approving all changes to the environment from a security perspective.

d. Privacy Officer from every agency

The Privacy Officer (PO) is responsible for evaluating and approving all changes to the environment from the perspective of supporting the State's privacy requirements.

e. Chief Financial Officer or financial representative

The Chief Financial Officer (CFO) has the responsibility of understanding the cost involved with the security and privacy decisions that are made.

f. Office of Technology Services[19] representative

The Office of Technology Services (OTech) representatives would be responsible for providing and administering the required resources and implementing the final technical solution.

Since the participation of this Change Management Committee is identical to the security Committee laid out in the Security and Privacy Approach identified in Section 5.7.4, the participants can be the same.

The Change Management Committee should focus on the business impact of the changes that are being administered into the Shared Data Cloud and support the agencies in leveraging the shared data assets to the fullest.

5.5 Avoiding the Pitfalls

In spite of the importance and permanence of organizational change, most change initiatives fail to deliver the expected organizational benefits. This failure occurs for a number of reasons, and must be forever in the forefront and planning so that they do not creep into your ongoing project.

- poor executive sponsorship or senior management support
- poor project management skills
- political infighting and turf wars
- absence of a change champion or one that is too junior
- poorly defined objectives/goals

[19] OTech was formally known as Department of Technology Services (DTS).

- change team diverted to other projects

Failed organizational change initiatives leave in their wake cynical and burned out employees, making the next change objective even more difficult to accomplish. It should come as no surprise that the fear of managing change and its impacts is a leading cause of anxiety in managers.

Change management is an important process. Understanding an organization and its objectives, and matching the initiative to the organization's real needs (instead of adopting the latest fad) is the first step in making a change program successful. Beyond that, recognize that bringing about organizational change is fundamentally about changing people's behavior in certain desired ways. As is apparent from the above list of reasons for failure, lack of technical expertise is not the main impediment to successful change. Leadership and management skills, such as visioning, prioritizing, planning, providing feedback, and rewarding success, are key factors in any successful change initiative.

5.6 Data Governance

IBM describes data governance as "a quality control discipline for assessing, managing, using, improving, monitoring, maintaining, and protecting organizational information"[xviii]. Data Governance is an approach for making decisions, assigning accountability, identifying business rules and defining processes related to information itself. .

Data governance is needed throughout the enterprise, however, this section details the data governance strategy specifically with respect to secure data sharing across the State. Internal agency data governance needs are not addressed in this document, although the same approach could be taken.

5.6.1 Governance Committee

A data governance committee and/or team is needed to align business priorities, people, processes, and technology in order to set policies and procedures for a shared data strategy administration. Data governance is essential to manage transactions, information, and knowledge necessary to initiate and sustain the activities identified in the shared data strategy. Without a governing body in place, a statewide data sharing initiative is not possible.

The data governance committee will decide the rules behind how data is consolidated, who owns the data, the update rules for the data, as well as the auditing requirements. In addition, the committee is responsible for deciding on the best way to resolve the data conflicts that are bound to occur with more complex update rules.

5.6.2 Governance Model

One example structure for this governance committee is as follows:

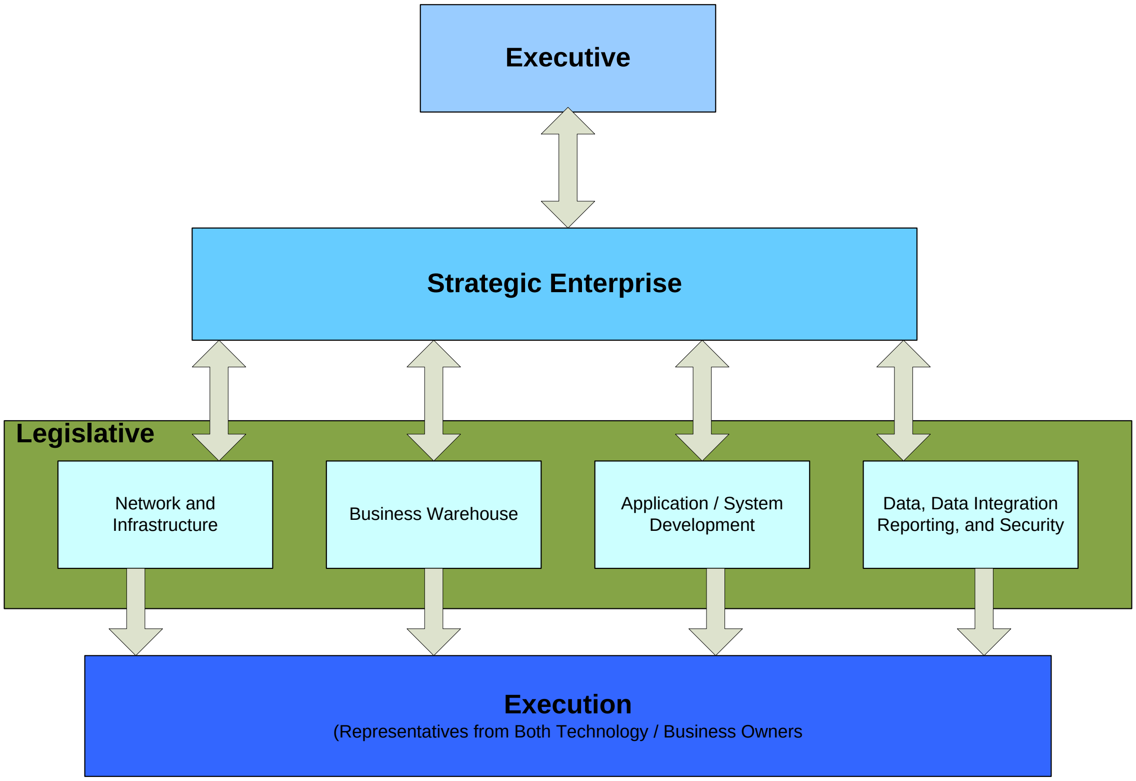

Executive
Strategic Enterprise
Legislative
Network and Infrastructure
Business Warehouse
Application / System Development
Data, Data Integration Reporting, and Security
Execution
(Representatives from Both Technology / Business Owners

Figure 5-2 – Governance Structure

5.6.3 Executive Level

The top of the governing level is the ***Executive*** level. The executive level of the infrastructure would incorporate the key operating strategies of sponsorship, strategic direction, funding, advocacy, and general governance oversight. Based on the current economic environment, the driver in the executive level will be the funding. Funding will need to be assessed from agency to agency to cross-agency considerations.

5.6.4 Strategic Enterprise Level

The next layer of the governing model is the ***Strategic Enterprise*** level. Managing and mediating any disagreements to the strategic planning actives fall under this level. This includes enforcement for the newly created policies and procedures. The intention of having this reside at this level is to ensure smooth management of the governance.

5.6.5 Legislative Level

At the ***Legislative*** level, senior business leaders from both business and technology commit resources to the governance team. As an option, the model suggests that you break out this level based on key technology functions, creating a type of stewardship. This stewardship would help in the focusing of key subcomponents of technology such as Network and Infrastructure, Business Warehousing, Application/System Development, Data, Data Integration, Reporting, and Security. The idea captured in having both business and technology present at this level is that they can commit resources to the governance. Committed resources will be paramount. If there are gaps in the policy design, this level must address and rectify these gaps. The established policies and procedures must be managed by the SLAs. This level ensures that the governance policies tie to the CIO's business strategies.

5.6.6 Execution Level

The final layer of the structure is the ***Execution*** level. Armed with the strategic directive developed at the legislative level, the execution layer's role focuses on implementing the identified strategic design and its directive. Embedded in this strategic design, the execution team has a long list of tasks to implement and monitor:

- Tracking of metrics
- Data Models
- Security
- Security Access
- Data Integrity
- Data Ownership
- Data Security
- Records Management
- Archiving
- Long term access to archives

Appropriate, effective, and secure data sharing cannot occur without a strong data governance model in place. The model above takes into account the alliance of people, processes, business, and technology, all moving toward the common goal of using shared data effectively. The creation of a permanent data governance advisory council would be a strong move forward in the area of statewide data strategy for the State of California.

5.6.7 Governance Areas of Focus for Data Sharing

Documenting, knowing, and understanding the data owned by the enterprise is essential in a data sharing initiative. As part of the initial charter for this Statewide strategy was to obtain and analyze the participating agencies' data. The analysis, review, and feedback from all the participating agencies revealed the need for special consideration for the following areas of focus.

5.6.7.1 Data Quality

The quality of the data being collected and maintained needs to be reliable and accurate. Reliable and accurate translates to data that is correct and precisely reflects the object or transaction that it is for, regardless of its origin. When an agency only acts as a pass through of information, say from the county, it is hard to ensure the data quality. Even though the agency from which the shared information is received may not be the owner, the trickledown effect is that many people attribute the quality of the data to system. Therefore, in the setup of the governance, key efforts will be made in establishing data ownership, validating data quality, and ensuring data timeliness.

5.6.7.2 Data Ownership

As stated in Section 8.2.1 (*Data Harmonization*), data ownership must be captured for each of the data elements. Since the data is shared, there will be multiple consumers of the data, but only a handful of potential owners. The ownership may change over a point in time. For example, the best source for a person's name and address may be from their tax records. However, the DMV may have updated information, but for only drivers and identification card holders. For people who do not pay taxes or drive, how will this data be captured – possibly from a system within CHHS? In this example, there were potentially at least three sources of information depending on the business rules. So data ownership may not be straightforward. The recommendation is to start simply, and evolve from there when the need occurs.

5.6.7.3 Auditing

The data, online views, and reports from the data need to be credible and certifiable to be usable. This means that the data must be able to be tracked from source to destination and verified for security and accuracy. Key considerations need to address the information from reports, and views should not be alterable.

5.6.7.4 Security

Securing the data is critical to the integrity of the data and the confidence of the agencies using the data. Policies and procedures will be addressed, developed, and enforced by the governance committee focusing on the assurance of proper handling. Security needs extend to extractions, data movements, loads, and reporting processes.

5.6.7.5 Secure Data Sharing Between Agencies

Memorandums of Understanding (MoUs) are detailed in the Office of Information Security (OIS) document State Information Management Manual (SIMM) 65E[20]. MoUs are used to define a relationship between two departments or two agencies. They create a platform for a clear understanding of each agency's commitment and expectations. Having them in place ensures smooth and secure data sharing between agencies. MoUs should include:

- Supersession – does this supersede another agreement?

[20] At the writing of SIMM 65E the Office of Information Security (OIS) and the Office of Privacy Protection (OPP) were once one office, OISPP.

- Introduction and purpose of the document
- The entity that authorized the data sharing (e.g., Legislature)
- Background information used to describe the systems and how they are connected
- The names and contact information of each party
- Rules of behavior. When will each party have access to the systems?
- The validity period
- What happens at the end of the validity period
- Any set dates to review activity, performance, or satisfaction
- What parts are open to change or negotiation and how
- What aspects should require formal notification and how
- Any restrictions
- Any disclaimer statements
- Any privacy statements
- Service level agreements (SLAs), penalties for not meeting the agreement
- Security, which details the security arrangements that each party will have to abide by
- Cost considerations are details as well as the financial commitments
- Service disruption and recovery details in case anyone of the interfacing systems needs to be recovered
- Other considerations such as licensing fees for technology and/or data
- Signature authority, where the representatives for each party will sign the agreement

MoUs will be developed and enforced by the governance committee. A sample MoU has been included in the report in Appendix G.

For the technical details pertaining to the MoU, that information can be placed into another document called an Interconnection Security Agreement (ISA). This agreement has many more details regarding the interfacing systems, contact information for the personnel supporting the systems, and a schedule for the support. Information about how the data is shared, classified and how it is described. Legal restrictions are brought out in the ISA as well, since not all information can be freely shared.

For more details on MoU and ISA please refer to OISPP document SIMM65E.

5.6.7.6 Budget

In the governance and the layout of data sharing strategy as a future project, budget plays both a direct and indirect role. Indirectly, all participating agencies have identified their shortcoming from a staffing and resources perspective, with the root of this shortcoming being the lack of funding. Directly, new projects need funding to begin, thus the Statewide data sharing project needs to establish firm commitments in funding from the very start. The State should start with a small but qualified staff for this effort. This small committed group could then begin to identify what cross funding is needed to engage each key agency in participation, segueing into establishing a proposal for budget.

5.6.8 Resources

Resources for the governance are assigned and addressed at the legislative level of the data governance model. Since the legislative level is composed of senior leaders from both business and technology, the authority, ability, and support to commit resources to the governance team should be widely embraced. As mentioned above, resources have a direct tie to budget and funding. Without the resources or budget to staff the governance, you do not have a key component to begin the state-wide data sharing project. The State should assign a business representative for the Data Governance Legislative body. All agencies should participate, especially those agencies using or planning to use the Shared Data Space.

5.6.8.1 Service Level Agreements and Service Level Objectives

Service-level Agreements, (SLAs), are an accepted standard in the IT and government industries. A service-level agreement is a negotiated agreement between two parties where one is the customer and the other is the service provider. As a shared data strategy has multiple parties and multi-level services, there will be many agreements that come into play as the initial data sharing infrastructure evolves. As discussed above, the data governance committee decides the rules behind how data is consolidated, owned, updated, and audited. These rules all need to be enforced, and the SLA is the correct tool to fortify the enforcement. Adding the elements of timeliness and completeness to the received data yields all the key factors that must be addressed in an SLA. Using the Master Data + Services Oriented Integration diagram, we pictorially depict where SLAs should exist.

A Service Level Objective (SLO) is many times used in government in place of a SLA. A SLA typically has damages associated with not meeting the agreement, and a SLO does not. In a government setting, the SLOs are typically used since enforcing damages on a government entity is simply not done. The problem with SLOs is that the consequence of not meeting an objective is many times not felt by the supporting organization. Caution with this approach should be given, as it hinders the adoption of shared services by the agencies. For example, CalEPA has SLAs that are in place with the federal government Environment Protection Agency. If those SLAs are not met then federal funds are withdrawn. If CalEPA is to use another department to provide these critical shared services, they will need to be assured that a certain level of service will be supported, and if the SLA is not met then the penalty will need to be shared by both parties.

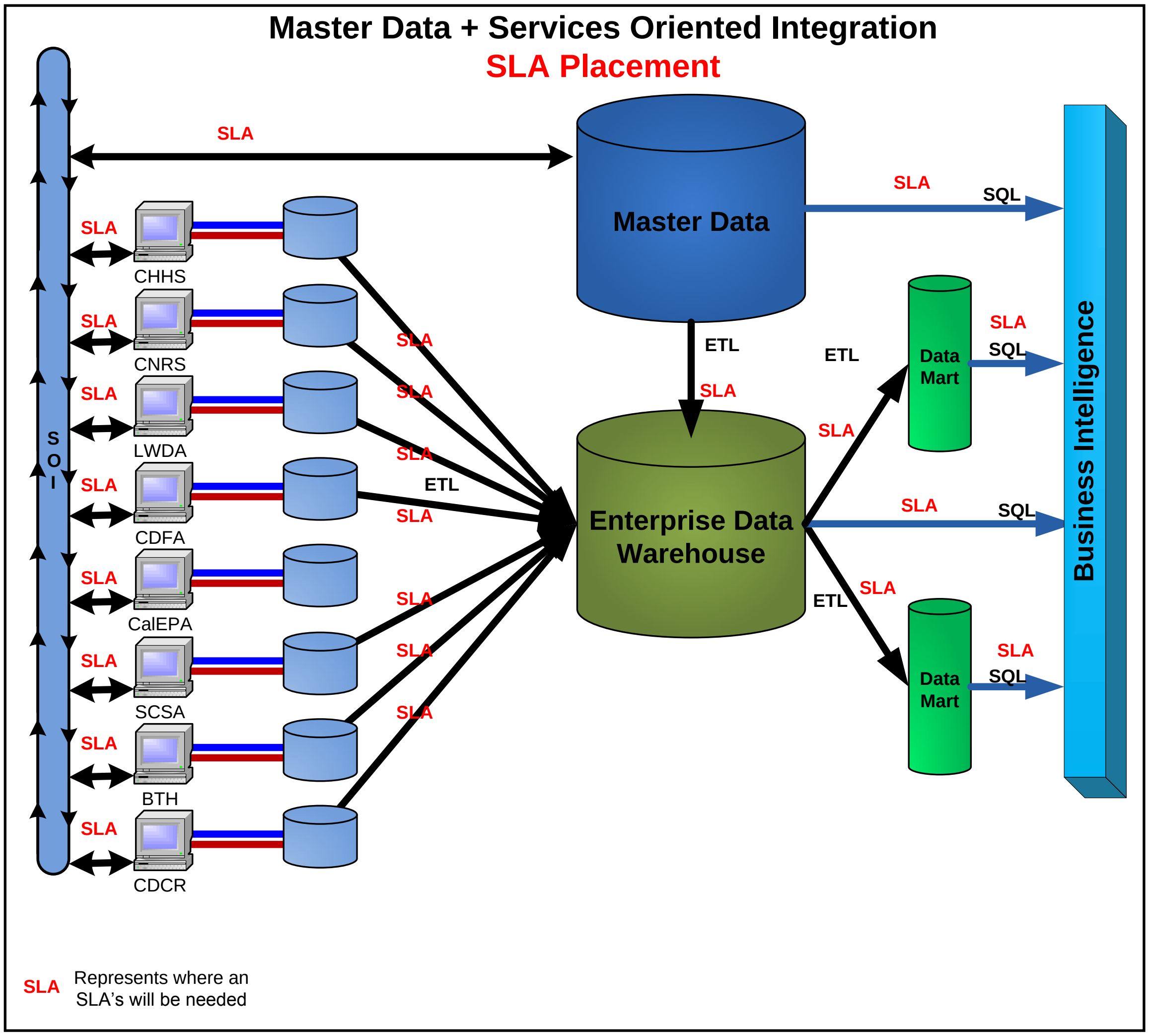

Master Data + Services Oriented Integration
SLA Placement
S O I
SLA
Master Data
CHHS
CNRS
LWDA
CDFA
CalEPA
SCSA
BTH
CDCR
ETL
Enterprise Data Warehouse
Data Mart
Data Mart
SQL
Business Intelligence
SLA Represents where an SLA's will be needed

Figure 5-3 – Service Level Agreements

5.6.8.2 Conflict Resolution

What happens when two agencies are trying to update the same record? In theory, this should not occur. The metadata associated with each transaction identifies who originated the transaction and in what context the transaction was generated. Therefore, identifying the data owner should be straightforward. With the data's business owner driving, the next step is to draft a policy and procedure around this ownership. When, by whom, and why would this data be updated? Security is also a consideration at this point. Are there any exceptions? When these questions have been exhausted and document, the next step is an SLA to manage failures to follow the rules.

5.7 Governance Execution

The governance process begins with setting objectives for the State – the OCIO provides the initial direction. From then on, a continuous loop is established. Performance is measured and compared to objectives, resulting in redirection of activities where necessary and change of objectives where appropriate. While objectives are primarily the responsibility of the Committee and performance measures that of management, it is evident they should be developed in concert so that the objectives are achievable and the measures represent the objectives correctly.

5.7.1 Governance Process Model

In response to the direction received, the IT function needs to focus on realizing benefits by increasing automation (making the enterprise more effective) and decreasing cost (making the enterprise more efficient) and on managing risks (security, reliability, and compliance). The IT governance framework then can be completed as indicated below:

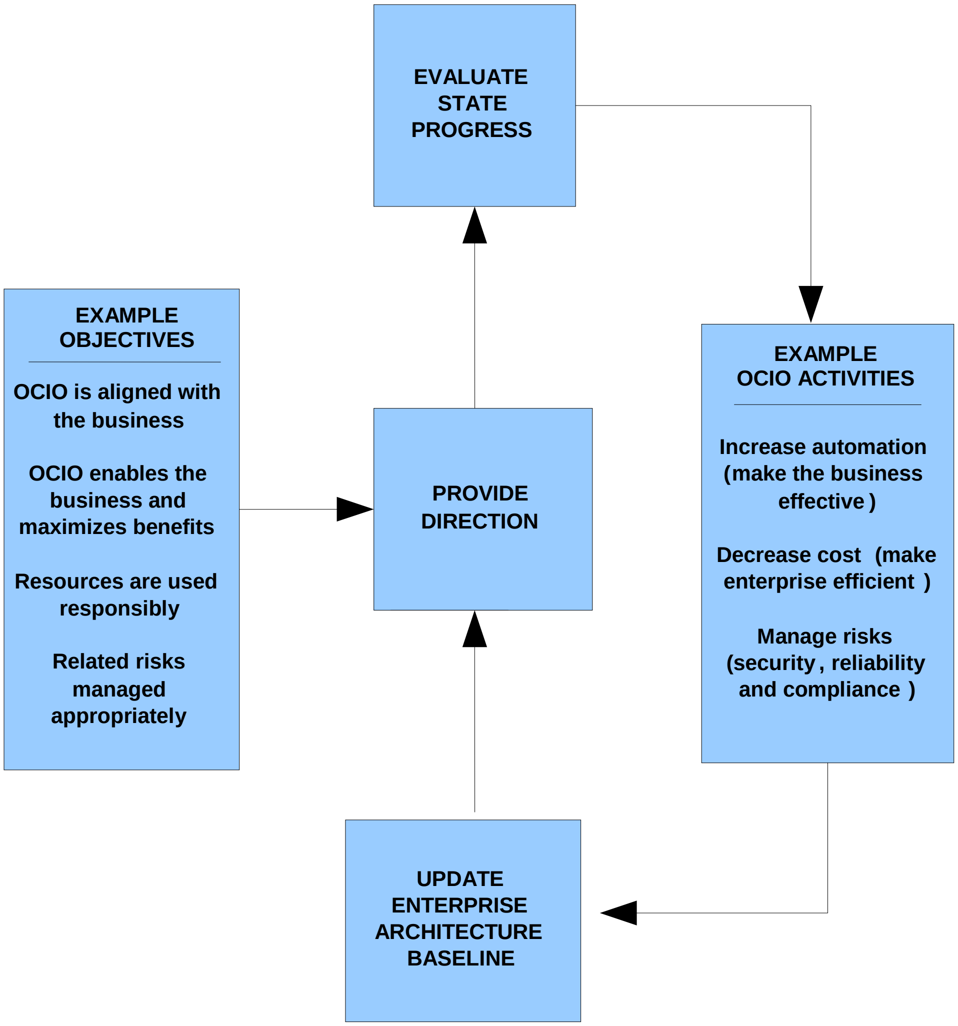

EVALUATE STATE PROGRESS
EXAMPLE OBJECTIVES
OCIO is aligned with the business
OCIO enables the business and maximizes benefits
Resources are used responsibly
Related risks managed appropriately
PROVIDE DIRECTION
EXAMPLE OCIO ACTIVITIES
Increase automation (make the business effective)
Decrease cost (make enterprise efficient)
Manage risks (security , reliability and compliance)
UPDATE ENTERPRISE ARCHITECTURE BASELINE

Figure 5-4 – Governance Process Model

Data is a State asset and resource. Data governance is really about properly managing the State’s data assets, information, and knowledge. Managing the expectations of how shared data is to be managed from the start through governance is essential to establishing a scalable statewide data sharing solution.

5.7.2 Getting Started

All the business processes offered by various State agencies have been organized according to a three level hierarchy established in the California BRM. The second level of organization is Community of Interest (COI). COI is the inclusive term used to describe collaborative groups of users who must exchange information in pursuit of their shared goals, interests, missions, or business processes and who therefore must have shared vocabulary for the information they exchange. Communities provide an organization and maintenance construct for data such that data goals are realized. Moving these responsibilities to a COI level reduces the coordination effort as compared to managing every data element department-wide. For example, standardization and control of data elements, similar to the current data administration approach, can be done at the community level rather than requiring all data elements to be standardized across the State.

Communities will form in a variety of ways and may be composed of members from one or more functions and organizations as needed to develop the shared mission vocabulary. In some cases a ‘community’ may have authority from explicit chartering. Institutional COIs, whether functional or cross- functional, tend to be continuing entities with responsibilities for ongoing operations. They also lend support to contingency and crisis operations. Expedient COIs are more transitory and ad hoc, focusing on contingency and crisis operations.

The COIs support users across the Enterprise by promoting data posting, establishing “shared” space, and creating metadata catalogs. Data within a COI can be “exposed” within the COI or across the State by having users and applications “advertise” their data assets by cataloging the associated metadata. These catalogs, which describe the data assets that are available, are made visible and accessible for users and applications to search and pull data as needed.

Although many of the COI functions will be similar regardless of COI characteristics, there will be some additional roles for institutional COIs. Institutional community members will collaborate to ensure that the necessary structures are in place to achieve the data goals. In particular, during the transition to net-centricity, institutional community members must take the lead in establishing COI-specific metadata structures, defining community ontologies, cataloging data and metadata, and having members post data. The COI-specific metadata structures provide an extended level of data definitions and structures, and the community ontology provides the data categorization, thesaurus, key words, and/or taxonomy. The COI-specific metadata structures and the community ontology serve to increase semantic understanding and interoperability of the community data. These community ontologies and data structures are visible to the Enterprise—by increasing visibility, data “stovepipes” will be mitigated.

The institutional COI efforts may enable the expedient COIs to quickly become operational when needed. The users in an expedient COI not only pull and use data but also create and post data to the Enterprise. A member of an expedient COI may leverage the data structures defined by the institutional COIs. For example, when providing metadata for a new data posting, the member can provide the metadata already defined in one of the institutional COIs’ schemas. However, expedient COIs can also create independent metadata structures, ontologies, and catalogs.

Based on the diversity of COI characteristics and roles, there will be a variety of operating processes and procedures that will be used by COIs to accomplish their data activities. Pilot activities with "trial COIs" will further refine the construct. More detail on COI functions will be provided in subsequent transition planning guidance.

5.7.3 Governance Maturity Model

As an organization embraces data governance and matures in governing their data, the over risk associated with poor data quality goes down and the benefit to the organization goes up. Figure 6-5 illustrates this benefit.

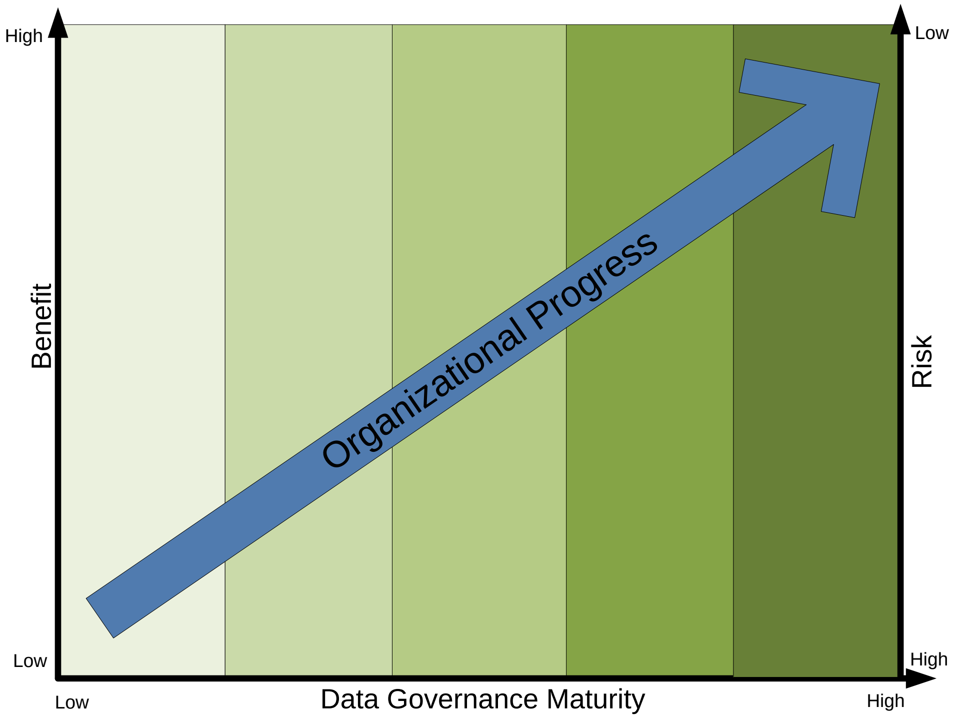


Figure 5-5 - Data Governance Maturity Model

An excellent discussion of the different maturity models that are available can be found in the National Association of State Chief Information Officers (NASCIO) Document 'Data Governance Part II: Maturity Models – A path to Progress' included in Appendix J.

The maturity models listed in the NASCIO document are very similar. Some identify 4 levels of progress (e.g., Oracle, MDM Institute, DataFlux) while others identify more levels (e.g., IBM, Knowledge Logistics, Gartner). Figure 5-6 is a compilation of these maturity models generalized for this discussion. The left bottom corner of the figure represents business environments with few governance business processes, poor quality data, and data duplication. These organizations find themselves with fragile, rigid data sharing solutions, and their ability to respond to business needs quickly is difficult. As you progress to the upper right corner, governance business processes are put into place, data duplication is removed, and data quality is addressed. The organization's ability to respond to changes in business improves. As the organization progresses into maturity, then the opportunity for proactive management and shared common data services increases. Benefit to the business increases while overall risk decreases.

5.7.4 Security and Privacy

As specified in the FEA DRM, security and privacy considerations apply to all three of the DRM's standardization areas. Data described, contextualized, and shared may include personal information and/or proprietary information that will trigger security and privacy requirements. For example, data sharing involving social security numbers may require chain of trust agreements. The Federal Chief Information Officers Council has created and is maintaining the Federal Enterprise Architecture Security and Privacy Profile (FEA SPP). It is a guide to promote best practices and recommendations for layers of security and privacy in enterprise architecture. The FEA SPP is a scalable and repeatable methodology to address information security and privacy requirements from a business-centric enterprise perspective. It enables end-to-end planning and coordination of efforts to implement security and privacy across all FEA reference models. To support enterprise architecture, the FEA SPP methodology:

- Promotes an understanding of an organization's security and privacy requirements, its capability to meet those requirements, and the risks to its business associated with failures to meet requirements.
- Helps program executives select the best solutions for meeting requirements and improving current capabilities, leveraging standards and services that are common to the enterprise or the federal government as appropriate.
- Improves agencies' processes for incorporating privacy and security into major investments and selecting solutions most in keeping with enterprise needs.

The FEA SPP methodology is composed of three stages. They are:

a. **Identification -** The goal for this stage is to fully identify security and privacy requirements that are applicable to a business process (defined as a sub-function of a COI in CalBRM). An Enterprise Architect with the assistance of an Information Security Officer would determine the requirements by first understanding various State and federal laws, policies of participating agencies, regulations, market practices, contracts, and performance objectives that are applicable to a business process. Once these requirements are determined, it would be necessary for the EAs to familiarize themselves with the current and planned capabilities of the system.

b. **Analysis -** In this stage, Enterprise Architects perform a gap analysis. They would be required to identify the gaps between requirements and current or planned capabilities. Once done, a proposal is presented in the form of one or more solutions to address the gaps or towards improving existing capabilities.

c. **Selection -** The final stage involves presenting outputs from both "Identification" and "Analysis" stages to the governing committee for the selection of a solution to implement.

5.7.4.1 Approach

In order to apply the methodology and recommendations that are documented in the FEA SPP, the first step is to establish a security and privacy committee. It would comprise:

a. **Business Executive from every agency**

The role of Business Executive is to provide input towards laws, policies, regulations, contracts, and performance objectives.

b. **Enterprise Architect from every agency**

The Enterprise Architects would be responsible for decomposing business processes and recommending solutions.

c. **Information Security Officer from every agency**

The Information Security Officer (ISO) is responsible for evaluating and approving all changes to the environment and all security recommendations.

d. **Privacy Officer from every agency**

The Privacy Officer (PO) is responsible for evaluating and approving all changes to the environment from the perspective of supporting the State's privacy requirements.

e. **Chief Financial Officer or financial representative**

The Chief Financial Officer (CFO) has the responsibility of understanding the cost involved with the security and privacy decisions that are made.

f. **Office of Technology Services[21] representative**

The OTech representatives will be responsible for providing and administering the required resources.

5.7.4.1.1 Process

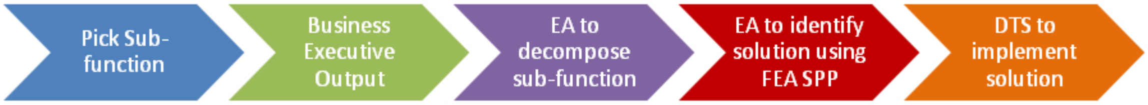

Pick Sub-function
Business Executive Output
EA to decompose sub-function
EA to identify solution using FEA SPP
DTS to implement solution

Figure 5-6 – Security Business Process

a. The committee would choose a sub-function of a COI that is defined in the CalBRM.

[21] OTech was formally known as Department of Technology Services (DTS).

b. The business executives would provide the following information:

    - Externally driven laws, regulations, and executive branch policies
    - Internally driven policies, interagency agreements, contracts, market practices, and organizational preferences
    - Mission-centric drivers such as performance objectives and lines of business

c. The Enterprise Architects along with the ISO and the PO would use the output from business executives to decompose the business process in order to understand and identify security and privacy requirements. They would then follow FEA SPP methodology to select a solution.

d. The CFO or financial representative will work with the Enterprise Architects to quantify the value and the cost of the solution.

e. OTech would provide the manpower and expertise to implement the solution. Since OTech will be executing the security plan, it is important that they be represented during the security process.

The committee would then move on to the next sub-function and follow the process again. This is an iterative process, and when one sub-function is complete, they would repeat the process until security and privacy is addressed for all sub-functions in CalBRM.

5.8 Office of Technology Services

Office of Technology Services (OTech), formally known as the Department of Technology Services within the State and Consumer Affairs Agency (SCSA), has been the department to support shared resources within the State Government. Changes will be needed to support this enterprise solution for the Shared Data Space. Some areas that will be operationally new to OTech are:

- Support of 24x7 operations
- Concept of SLAs as opposed to SLOs
- Dedicated and Trained staff to the Shared Data Space

5.8.1 24x7 Operations

Currently OTech does not technically support a 24x7 shop. The after hour's staff that is on site typically cannot cover most technical issues. Their main purpose is to escalate issues to technical staff to fix the problem. These technical resources are paged, and must stop what they are doing to address the problem. This has worked well in the past, as the IT systems under OTech control can sustain moderate downtime without a major disruption to the business.

Once critical departments like the California Highway Patrol, CalFire, and California Department of Corrections are participating in the data sharing, higher availability may be required. To sustain this higher availability, technical staff may be needed on-site to address any issues that may be encountered.

5.8.2 Service Level Agreements vs. Service Level Objectives

Service Level Agreements (SLA) and Service Level Objects (SLO) are discussed in detail in Section 5.6.8.1. The main difference between the two is that SLAs have a penalty and SLOs do not. The concern is the perception that the SLOs do not encourage the same sense of urgency as a SLA and therefore the associated quality of support is much lower. This concern was raised in our meetings with the agencies and needs to be addressed. In addition, if a service level penalty exists for an agency that is supported internally, all should share in the penalty if the service level is not met. If this is not done, the agencies will lose confidence that their external SLAs will be taken seriously.

5.8.3 Staff and Training

Additional training and staff will be needed by OTech for the management of the Shared Data Cloud. Even though OTech manages a diverse technological landscape, some of the technology identified to support the Shared Data Cloud is new, and therefore the staff will need to be trained. In addition, a dedicated staff should be considered to support the environment and the committees for data governance and business change management.

5.9 Summary

The organizational changes necessary to support the strategy fall in two categories, change management and governance. Change management is about managing changes to the business, the ongoing projects, Shared Data Space itself, and on-boarding new trading partners. Governance is about managing the data that is shared, understanding the business rules behind the data, ownership of the data, security requirements, and the expected lifecycle of the data. In addition, the governance committee will identify the service levels necessary to support the data. To enable them, business processes have been identified and recommended.

This page intentionally left blank

6. Work Plan

This section provides a work plan for each of the project phases that have been identified, phase durations, dependencies, anticipated resources (staff, software, and equipment), and a high level cost estimate. The strategy is vendor independent, start dates are not set and the resources have not been identified, therefore only high level information on costs, schedules, and resources is provided. In addition, it is recognized within the State the procurement process for hardware, software, and human resources can take many months or even years. These unknowns cannot be fully addressed in the work plan estimates.

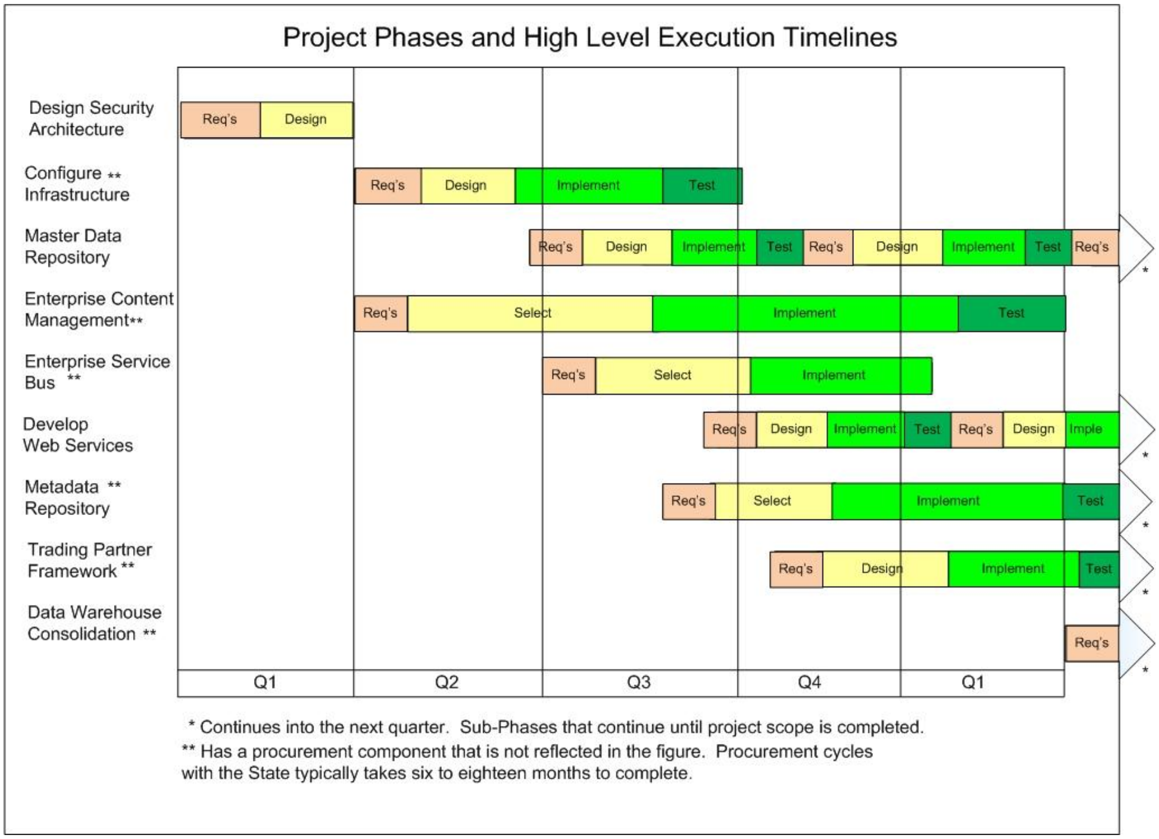
6.1 Overview

The work plan is a summary of the project plan templates provided in Appendix A. The goal of the work plan is to provide a high level overview of each of the eight phases, providing details in the following areas:

- Phase name
- Description
- Duration
- Project Management Methodology
- People Needed
- Hardware/Software Needed
- Dependencies
- *Estimated* Rough Order of Magnitude (ROM) Cost

The details that are identified for the project should be considered as a template, and a starting point for a project schedule moving forward. Detailed tasks and the associated resource assignments are included in Appendix A for each of the project phases. In addition, an overarching plan is provided in Appendix A to document the dependencies between the phases.

6.2 *Phases*


Project Phases and High Level Execution Timelines

* Continues into the next quarter. Sub-Phases that continue until project scope is completed.
** Has a procurement component that is not reflected in the figure. Procurement cycles with the State typically takes six to eighteen months to complete.

Figure 6-1 - High Level Execution Timeline

There are nine major project phases that have been identified in chapter 4 and are illustrated in Figure 6-1. This figure assumes that the concept, FSR and BCP has already been submitted and approved for the overall project. The major project phases are:

- Security Infrastructure
- Configure Infrastructure
- Build Master Data Repository
- Select Enterprise Content Management
- Select Enterprise Service Bus
- Develop Web Services
- Build Metadata Repository
- Build Trading Partner Network
- Consolidate Data Warehouses

Either a waterfall or spiral project methodology will be used in the execution each of the phases. The phases are interrelated as together they complete the data strategy and they build upon themselves. The interdependencies are identified within the work plan.

To provide more detail for each of these project phases a high level work breakdown structure has been provided in Appendix A.

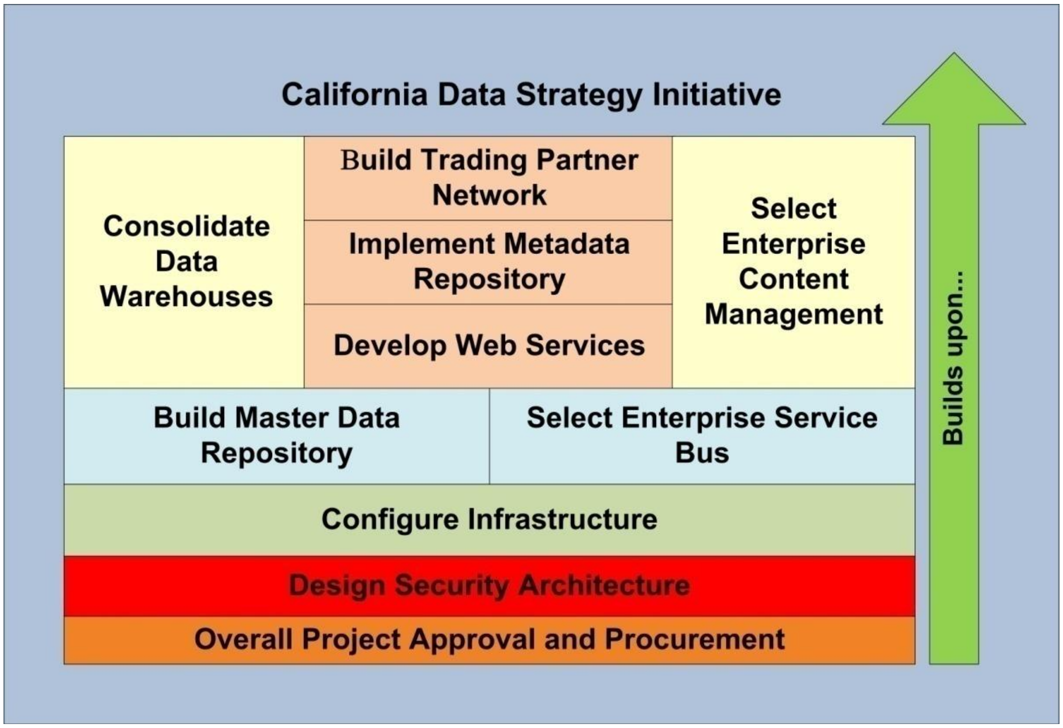

California Data Strategy Initiative
Consolidate Data Warehouses
Build Trading Partner Network
Implement Metadata Repository
Develop Web Services
Select Enterprise Content Management
Build Master Data Repository
Select Enterprise Service Bus
Configure Infrastructure
Design Security Architecture
Overall Project Approval and Procurement
Builds upon...

Figure 6-2 – CDS Phase Dependencies

Figure 6-2 describes the major dependencies between the project phases. The Security Architecture is the foundation for CDS and within each phase security work has been identified. The physical environment is created in the Configure Infrastructure phase. The Master Data Repository and the Enterprise Service Bus become the two main features of CDS which support the remaining phases.

6.2.1 Estimate Methodology

The phases of the data strategy work plan are identified at a high level. These work breakdown structures, in the form of a high level 'project plans', identify resource assignments and dependencies. Resourcing, estimated cost and overall durations have been identified from these plans. A work breakdown structure view of the project plans have been provided in Appendix A.

The following assumptions have been made for each of the project phases:

- Estimates of both time and cost are based on the assumption that the approval cycle for the concept, FSR and BCP will be limited to 6 months.
- Knowledgeable subject matter experts will be available when needed.

6.2.1.1 Cost Estimate

The cost for each of these phases is a Rough Order of Magnitude (ROM) estimate intended to give a general idea of the cost. Since the strategy is vendor independent, the timeline is not set and the skill level of the resources is not identified, the cost should be taken only as an estimate. At the onset of each phase a more precise estimate of the cost and a detailed timeline should be prepared.

The resource cost has been calculated as $100/hr blended rate for all resources. The timeline estimate was based on high level project plans created from the analysis. The resource cost was taken from these project plans.

The software cost was calculated as list cost minus the standard[22] discount for the State. Select vendors were evaluated. It should be noted, that the final software cost is usually much lower than the list cost with the standard discount. Vendors have been supplying the state with enterprise and unlimited license agreements that drive this final cost to the State much lower. These agreements are calculated from specific input which is not available at the time of writing this strategy.

The cost range calculation is based on the estimated minimum configuration cost and the estimated maximum configuration cost. Rounding of the final result is used to calculate a minimum and a maximum cost.

6.2.1.2 Resource Estimate

The resource assignments in the work breakdown structures identified in ***Appendix A*** are at the role level. Multiple people may be required to fill a single role. Since staffing has not been addressed the resource assignments should be considered **Full Time Equivalent (FTE)** resources. The resource calculations are based on the manpower estimate for each role divided by the duration of the project. Resources requirements are rounded up. If the resource is needed over 50% of the time then one FTE has been identified. If the resource is needed less than 50% of the time then the resource is identified as “part-time”.

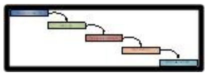
Procurement resources are not identified in the work plans.

6.2.2 Data Strategy Overall Project Approval and Procurement

This work plan manages the overall interdependencies between the different work plan phases. In addition, it also has the work required for the initial concept, Feasibility Study Report (FSR) and the accompanying Budget Change Proposal (BCP).

Duration: Six Months Execution[23]

Project Management Methodology: Waterfall



[22] With some vendors the standard discount is 50%.

[23] The State procurement process can take up to 18 months to complete.

People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	Business Analyst

Hardware/Software Needed: None

Dependencies: None

Assumptions:

1. Subject matter experts and technical experts are available to assist with the concept, FSR and BCP.

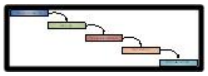
Estimated Cost: $300,000 to $400,000

6.2.3 Design Security Architecture Phase

This phase identifies and designs the security architecture used by the whole project. A comprehensive approach must be taken to ensure a secure solution. Designing security into the CDS ‘after the fact’ is simply not an option. The work that goes into this phase will identify the security architecture with respect to the network, the servers and the access to the system. All project phase and each product selected should use a similar approach to security so that the overall architecture can be secured consistently. Security across the State is constantly improving and this work is to ensure alignment with the State’s overall plan.

Duration: Three Months Execution

Project Management Methodology: Waterfall



People Needed:

FTE Quantity	Role
1	Project Manager (part-time)
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer[24]
1	Database Administrator (part-time)
1	Server Administrator (part-time)
1	Network Engineer (part-time)

[24] One may be required from every participating agency.

Hardware/Software Needed: None

Dependencies: None

Assumptions:

1. Identity management for the State will be identified.
2. Security procedures are in place for the targeted data center.
3. Subject matter experts and technical experts are available to provide the security input needed for each phase of the project.

Estimated Cost: $400,000 to $440,000

6.2.4 Infrastructure Configuration Phase

This phase builds out the actual infrastructure for the Shared Data Space discussed in Section 3. The following environments are supported:

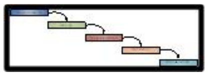
- Development
- Test
- Staging
- Production

The *production* and *staging* environments are scaled the same and are highly available. The development and test environment are scaled according to the need.

Until vendors are selected the cost is only a high level estimate. For estimate purposes, the hardware identified in this section is considering servers that are commodity hardware with dual quad core Intel processors running Linux. Larger more powerful servers can be used instead of commodity servers. However, the final cost must be calculated after the hardware vendor is selected.

Duration: One Year Procurement
Six Months Execution

Project Management Methodology: Waterfall



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Database Administrator
4	Server Administrator
1	Network Engineer (part-time)

Hardware/Software Needed:

Quantity (by server)	Hardware/ Software	Product
10	Hardware	Database server
8	Hardware	Application server
8	Hardware	Web server
2-6[25]	Hardware	Load balancers
2	Hardware	SAN Hubs
2	Hardware	Storage SAN
1	Hardware	Monitoring Server
27	Software	OS
10	Software	RDBMS. Examples include; DB2 Integrated Cluster Environment (ICE) or Oracle Real Application Clusters (RAC)
8	Software	Application server
8	Software	Web server
1	Software	Monitoring SW (e.g., Oracle Grid Control)

Dependencies: Nothing

Assumptions:

1. A clustered database will be used.
2. Vendor selection and procurement of hardware and software limited to 1 year.
3. Vendor selection for all components has been made.
4. Servers will be Intel on Linux.

Estimated Cost: $4.3 Million to $5.3 Million

[25] The number of load balancers used is dependent on the type used.

6.2.5 Master Data Repository Phase

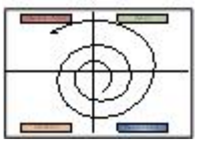
The master data repository will initially contain address information. The starting point for this phase should be the fields identified in the Federal Geographic Data Committee (FGDC) Street Address Data Standard (included in Appendix H). From there the following subject areas will be instantiated into the Shared Data Space:

- Organization (Businesses)
- Facilities
- Projects
- Services Offered
- Public record data deemed shareable
- Person (Constituents)

Analysis will be required at the beginning of each of these phases to determine the details of the data that will be shareable. The order can change but due to privacy issues ***person*** is recommended to be incorporated last.

Duration: One Year Procurement
Six Month Cycles

Project Management Methodology: Spiral



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Business Subject Matter Experts
1	Data and Process Modeler
1	DBA (Part-time)
1	Data Architect
1	Data Warehouse Architect
2	Testers (To validate migrated data)

Hardware/Software Needed:

Quantity	Hardware/ Software	Product
6[26]	Software	Database modeling software (e.g., Erwin Data and Process Modeler).

[26] Software quantity is dependent on how many teams are actively modeling their business processes. For pricing purposes number of agencies participating was set at six.

Dependencies: Design can occur anytime but implementation and or construction of the solution must be done after the Shared Data Space infrastructure is in place.

Assumptions:

1. Agencies will provide subject matter expertise to support this phase.
2. Six Agencies[26] will be actively involved.
3. Procurement of data and process modeling tool conducted a head of time.

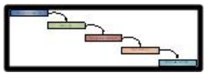
Estimated Cost: $780,000 to $1,100,000 Initially
~$375,000 per additional spiral

6.2.6 Metadata Registry Phase

The metadata registry is an integral component of the overall solution. The initial phase will be a single six month project phase to determine the best third party metadata registry available. The metadata registry catalogs all of the data assets and makes them discoverable to the business. As the overall project progresses and more services are added to CDS, the metadata registry may need minor changes to its configuration. These minor changes are not reflected in the work plan.

Duration: Six Month

Project Management Methodology: Waterfall



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Business Subject Matter Experts
1	Enterprise Architect
1	Data Architect
1	Testers
1	System Administrator

Hardware/Software Needed:

Quantity	Hardware/ Software	Product
Unknown[27]	Software	Metadata Registry

Dependencies:

The implementation must be done after the Shared Data Space infrastructure is in place. The Master Data Repository and any Web Service must be in place prior to their registry in the Metadata Repository.

Assumptions:

1. Agencies will provide subject matter expertise to support this phase.
2. Does not include seeding metadata registry with data.
3. Prior to registration the Master Data Repository must be built.

Estimated Cost:

$650,000 to $720,000

6.2.7 Enterprise Content Management Phase

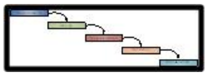
This phase is the selection of an Enterprise Content Management system. These products are mature and to purchase one is more desirable then building one. This phase is the selection of the Enterprise Content Management solution, its implementation and to test the configuration.

Duration:

One Year Procurement
One Year Implementation

Project Management Methodology:

Waterfall



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Business Subject Matter Experts
1	Database Administrator
1	QA Tester
1	QA Lead (Part–time)
1	Business Analyst (Part-time)

[27] Software quantity is dependent on how the software is priced. Pricing can vary by vendor.

Hardware/Software Needed:

Quantity (by server)	Hardware/ Software	Product
4[28]	Software	ECM Software

Dependencies: Infrastructure Configuration

Assumptions:

1. Approval for the concept, FSR, and BCP has been already received. A requirement for procurement to start.
2. One year timeline does not include the procurement time.
3. Structure for document metadata and records management disposition policies rules will be identified prior to implementation.

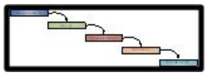
Estimated Cost: $2,000,000 to $2,800,000

6.2.8 Enterprise Service Bus Phase

This phase implements an enterprise service bus. Section 8 of the strategy addresses the standards that the Enterprise Service Bus (EBS) must support. This phase is to select, configure, test, and deploy the Enterprise Service Bus.

Duration:
One Year for Procurement
One Year Implementation

Project Management Methodology: Waterfall



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Database Administrator (Part–time)
1	System Administrator (Part–time)
1	QA Tester
1	QA Lead (Part–time)
1	Business Analyst (Part-time)

[28] Software quantity is dependent on how the software is priced. Pricing can vary by vendor.

Hardware/Software Needed:

Quantity (by server)	Hardware/ Software	Product
8	Software	EBS Software

Dependencies: Infrastructure Configuration

Assumptions:

1. Approval for the concept, FSR, and BCP has been already received. A requirement for procurement to start.

Estimated Cost: $1,600,000 to $1,700,000

6.2.9 Web Services Phase

This service will enable retrieval of address information for a given point in space. The following environments are supported:

- Development
- Test
- Staging
- Production

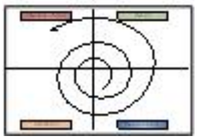
The production and staging environments are scaled the same and are highly available. The development and test environment are scaled according to the need.

The following web services will be built. Each web service will be built in a spiral development cycle. Each cycle is approximately 6 months in duration from requirements gathering to final testing and deployment. Multiple services may be written simultaneously depending on the final detailed schedule that is produced.

Web Service	Description
setAddress	Creates or updates an address record.
getPointFromAddress	Pass a standard address, datum ellipse and projection model into it and it will validate the input and return an x value, y value. Datum and projection are optional.
getAddressFromPoint	Send it x value, y value with a datum ellipse and a projection and it retrieves the nearest address information associated with the address layer.
isAddressValid	It validates address and reformats it into a standard format.
isValidPoint	Validates x and y values against a datum and a projection model. Is it a valid point in California?
getGISLayer	When supplied a valid point it will retrieve one, multiple or all of the GIS framework layers that contain the point.
getDistance	Get the distance between two points, a point and a line or a point and a polygon.

Duration: Six Month Cycles

Project Management Methodology: Spiral



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Business Subject Matter Experts
1	Enterprise Architect (part-time)
1	Data Architect (part-time)
1	Developers
1	QA Tester
1	DBA (part-time)

Hardware/Software Needed:

Quantity	Hardware/ Software	Product
4	Software	Development tool to develop module for data source integration
1	Software	Test Automation Tools

Dependencies: Infrastructure Configuration, Enterprise Service Bus

Assumptions:

1. Agencies will provide subject matter expertise to support this phase.
2. Four developers will be needed at any given time.

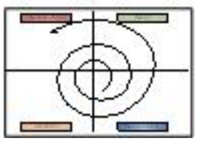
Estimate Cost: $ 1,300,000 to $1,400,000

6.2.10 Data Warehouse Phase

The data warehouse is predominately a dimensional view of the master data. Since the agencies already have their warehousing needs covered by stove-piped solutions, it is recommended to develop the enterprise data warehouse once sufficient master data has been instantiated into Shared Cloud to report on.

Duration: One Year Procurement
Six Month Cycles

Project Management Methodology: Spiral



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Business Subject Matter Experts
1	DBA (Part-time)
1	Data Architect (Part-time)
1	Data Warehouse Architect
2	QA Testers (To validate data and reports)

Hardware/Software Needed:

Quantity (by server)	Hardware/ Software	Product
4	Software	Business analytic or reporting software

Dependencies: Design can occur anytime but implementation and or construction of the solution should be done after several iterations of the master data repository and after the Shared Data Space infrastructure is in place.

Assumptions:

1. Agencies will provide subject matter expertise to support this phase.
2. Business Intelligence (BI) tool estimated on a per processor charge.
3. Procurement of BI tool is conducted prior to the start of the six month spirals.

Estimated Cost: $ 700,000 to $2,200,000[29]

6.2.11 Trading Partner Network Phase

This phase builds out the actual infrastructure for the trading partner network. The following environments are supported:

- Development
- Test

[29] The cost can vary widely depending on the analytic tool selected and the options within the tool that is chosen. There largest variability in the estimated cost is the software cost.

- Staging
- Production

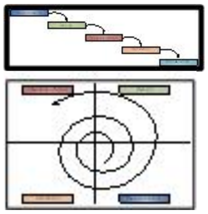
The production and staging environments are scaled the same and are highly available. The development and test environment are scaled according to the need. The externally facing web services that run on the trading partner network will be written once the standard internal web services are available. The infrastructure will be established by a waterfall project methodology however the web services can be created using a spiral methodology.

Duration:

One Year Procurement
Six Months Initial and 3 Month Spirals for the Externally facing services

Project Management Methodology:

Waterfall/Spiral



People Needed:

FTE Quantity	Role
1	Project Manager
1	IV&V (part-time)
1	IPOC (part-time)
1	Information Security Officer (part-time)
1	Server Administrator
1	Project Manager
1	Network Engineer (part-time)
1	Developer
1	QA Tester

Hardware/Software Needed:

Quantity	Hardware/ Software	Product
6	Hardware	Application server/Web server
6	Software	Application server
6	Software	Web server

Dependencies:

Infrastructure Configuration, Enterprise Service Bus and Web Services

Assumptions:

1. Hardware is procured and installed in the data center

2. Vendor selection for all components has been made
3. Agencies will provide subject matter expertise to support this phase
4. Security work for the DMZ environment is in place

Estimated Cost: $ 500,000 to $700,000 for initial spiral

6.3 Summary

The goal of the work plan is to address the phases in well defined sub-projects with a typical duration of 6 months to 1 year. Multi-year projects can easily lose their focus in the requirements and design phases of the project. Many times in the construction or implementation phase of a project is where it is identified that information was missed in the previous phases of the project. Shortening the project lifecycle into shorter sub-project phases ensures tangible and demonstrable progress. The technology used in the Shared Data Space is well suited for this type of project approach.

	Hardware Costs		Software Costs			Rounded Up Total Cost	
Initiative	**Est. Max**	**Est. Min**	**Est. Max**	**Est. Min**	**Estimated Resource Costs**[30]	**Est. Max**	**Est. Min**
Overall Project					$322,400	$400,000	$300,000
Design Security Architecture	$0	$0	$0	$0	$439,200	$440,000	$400,000
Configure Infrastructure	$2,113,000	$1,261,000	$2,246,750	$2,246,750	$868,800	$5,300,000	$4,300,000
Master Data Repository	$0	$0	$300,000	$60,000	$721,600	$1,100,000	$780,000
Metadata Registry	$0	$0	$120,000	$60,000	$591,600	$720,000	$650,000
Enterprise Content Management	$0	$0	$2,070,000	$1,380,000	$690,000	$2,800,000	$2,000,000
Enterprise Service Bus	$0	$0	$912,000	$912,000	$690,000	$1,700,000	$1,600,000
Web Services	$0	$0	$43,000	$0	$1,355,648	$1,400,000	$1,300,000
Data Warehouse	$0	$0	$1,776,000	$300,000	$411,200	$2,200,000	$700,000
Trading Partner Network	$210,000	$90,000	$150,000	$150,000	$318,400	$700,000	$500,000
Total	$ 2,323,000	$ 1,351,000	$ 7,617,750	$5,108,750	$6,408,848	$16,760,000	$12,530,000

In Appendix A, work breakdown structures for all of the phases are listed, detailing the task names, prerequisites and the type of resource assigned.

[30] Resource costs do not reflect Procurement resources and Subject Matter Expert resources required for creating the concept, FSR and BCP.

7. Risk/Issues and their Mitigation

This section will identify risks, issues and their corresponding mitigation strategies towards implementing Statewide Data Strategy.

7.1 Business Change Management

As with any initiative, the real challenge is almost always aligning the business with the final solution. One should not underestimate the risks associated with changing business processes and the project team should have a good understanding of the overall impact the project has on the business.

7.1.1 Risk/Issues

Some risks and issues that exist with the business are:

- Due to a lack of trust data is not shared with the other agencies
- Agencies not understanding the value proposition of validating and sharing data
- Political roadblocks between agencies making sharing difficult.
- Concern that the infrastructure cannot support the agency availability requirements.
- OTech must transition from Service Level Objectives to a model that supports Service Level Agreements.
- Subject matter experts must be available that are knowledgeable in the business area being worked.

7.1.2 Mitigation

Changing business processes is not only risky but also difficult as people have become accustomed to the current process and change is difficult. In addition, it is challenging to understand the implications associated with making a change to the process, especially if the business process is complex. To identify and encourage change, two committees (i.e., Change Management and Data Governance Committees) have been proposed to work through the issues. They will identify the change and work with the Agencies to understand the impact of the change to the business.

7.2 Funding

Funding any initiative with the current economic climate is not trivial. Since there are limited funds available within the State, only high priority projects are even considered. These funding challenges underscore the importance of understanding and communicating the value proposition of improved data sharing to the Agencies.

7.2.1 Risks/Issues

- In this rough economic climate with many States, including California, running a deficit, Funding may be a challenge.

- Even though data sharing will lower the cost for the State to do business, the initial cost will be higher during implementation in order to keep from impacting the current business processes.
- Coordination among the Agencies for the funding of the Shared Data Space may be difficult.

7.2.2 Mitigation

Understanding the cost or savings associated with all phases of this project is paramount. The problem is complex, and the solution challenging, as there are projects that are currently underway that overlap, compete with, and influence the data sharing strategy. The data governance committee and the change management Committee must understand the estimated benefit and cost to the business for each phase of the project. Prior to initiation of a phase, the Return on Investment (ROI) must be documented and well understood.

7.3 Agency Resources

Many agencies and departments are doing more work with fewer resources as budgets are tight and requirements that are placed on them increase. These organizations may not be able to supply the necessary resources to support the new initiatives.

7.3.1 Risks/Issues

- Agencies may not be able to supply knowledgeable subject matter experts to document the business adequately due to limited staffing and tight budgets.
- Agencies may provide non-dedicated resources in place of dedicated resources, thus impeding the ability to make decisions in a timely fashion.
- Agencies may assign the wrong staff to the project; ones that do not have the authority to make decisions or staff that are not decision makers.

7.3.2 Mitigation

It is again important to understand and communicate the value proposition to the Agencies. Agency should assign staff that is dedicated to the success to the Statewide Data Sharing strategy. Staffing should be consistent and rotation, at least initially, should be minimized. In addition, the staff assigned should have the authority to make decisions and should be regarded as decision makers.

7.4 Liability

When accurate data is consolidated and made available, then the risk of liability can increase from that data. Data that is sensitive in nature can create numerous concerns if breached. Security will need to be planned, implemented and monitored to ensure that the liability of sharing sensitive information is minimized. SIMM 65D-*Security Breach Involving Personal Information: Requirements and Decision-Making Criteria for State Agencies* identifies potential ‘harm’ to include but not limited to:

- harm to reputation,
- harassment,
- prejudice (particularly when health or financial benefits information is involved),
- financial loss,
- embarrassment,
- legal problems, and
- identity theft.

These risks are being managed today with the current IT infrastructure that is in place. However, as data is cleaned up and centralized, it becomes more valuable and causes a greater impacted in a security breach.

7.4.1 Risks/Issues

- Sensitive data could be made public by an agency causing harm to constituents.
- Data updates not timely and therefore decisions are made off of poor or inaccurate data.
- The master data repository could be a target for hackers since it is a central, accurate and comprehensive source of shared data.

7.4.2 Mitigation

Both sensitive data and data synchronization issues exist today and the State must rely upon the due diligence of each of the Agencies for risk mitigation. The processes in place are typically adequate to manage the liability issues associated with the data. Timeliness of updates is currently being managed independently among the Agencies with limited coordination. The data sharing strategy addresses these issues and more. The data governance committee, in conjunction with the data owners, will identify the *restrictions on disclosure and use* as well as security (e.g. confidentiality, integrity, and availability) requirements on each sharable piece of information. Then service level agreements will be created to ensure and track performance of these agreements. In addition, the ability to audit who requested which information will be available which will provide a greater level of security then what is in place today.

7.5 Privacy

Privacy is a big issue when it comes to personal information related to individual constituents. The goal of obtaining an accurate and comprehensive view of each of California's constituents can be realized by simply using and coordinating the information that is currently being captured and no more. Since the information is no longer fragmented across agencies and is now 'clean' and valid data, the new concern is with privacy of that information.

7.5.1 Risk/Issue

Now that the Agencies can ‘connect the dots’ on information about constituents, there could be major ‘pushback’ from the citizenry regarding the data sharing strategy. Concerns always exist with the government taking on the role of ‘big brother’. The Data Strategy only focuses on data that is currently available within the State, however, the fact the data will be harmonized and centralized may be enough to raise concerns.

7.5.2 Mitigation

To address privacy we must go back to our security and auditing rules for the information. The data sharing strategy including the shared data space will not contain any information that is not currently available within the Agencies. In addition, care will be taken with the data sharing agreements to ensure that the Agencies follow privacy and security protocols. Finally, the data sharing strategy focuses on sharing information that is less sensitive first so that the ‘kinks’ can be worked out before supporting more sensitive information. The State must provide notification of how the data is used to the State’s constituents and ensure that the privacy statutes, California State Statutes 1798, are maintained.

7.6 Subject Matter Experts

Good business requirements should originate from the business itself. The challenge always is finding the right experts who can identify and communicate these requirements. Budgetary challenges do not make this any easier. As resourcing becomes tight, department managers tend to offer up their newer resources as the subject matter experts for new initiatives. Without good requirements, a good solution is difficult.

7.6.1 Risk/Issue

Poor subject matter expertise can cause significant impact to the business as they may not know the business well enough to identify the issues, document the requirements or understand the benefits. Finding these experts is challenging as they are often in demand.

7.6.2 Mitigation

Escalate this need within the management of each department. Make providing subject matter experts a requirement for participation.

8. Concepts and Standards

This section identifies the industry concepts and standards needed to support the data sharing architecture and design discussed in Section 3. The following concepts are covered in this section; Data Harmonization, System Availability, Disaster Recover and Information Lifecycle Management. In addition, the standards that are used in the overall approach, like service oriented integration, web services, “Extraction Transformation and Load “, and Trading Partners, are discussed as well. It offers recommendations, guidelines and policies adoptable both State- and agency-wide to enable maximizing reuse of software and data.

8.1 Overview

Implementing any architecture involves a degree of risk related to the maturity/immaturity of the underlying industry best practices, key concepts, and standards. The State of California is not in the business of IT, but rather the business of government, servicing the constituents of the State and supporting State law. Therefore care was taken to select mature standards and business concepts to support the recommended California Data Strategy.

This section is organized into two parts: Concepts and Standards. A concept is an industry best practice that is used to support architecture similar to what is proposed in the Data Strategy Report. A standard is backed up by a standards body. The standards selected in this report are well known to commercial industry.

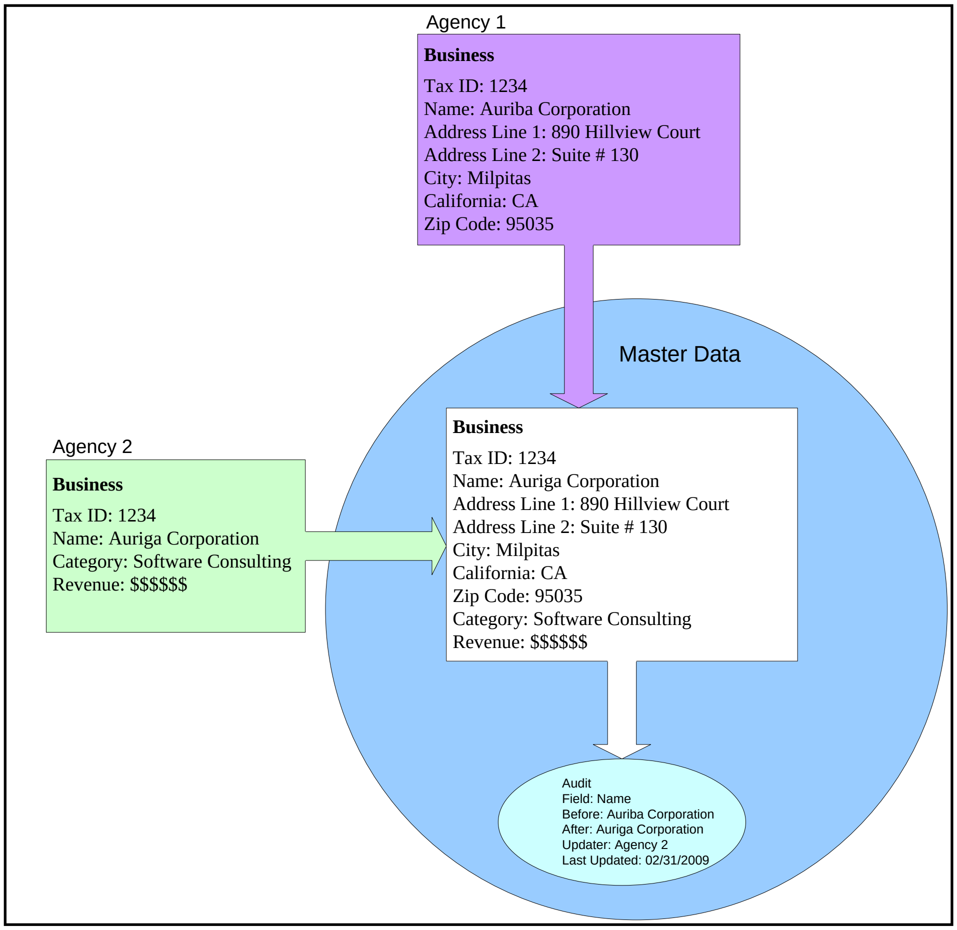
8.2 Concepts

The concepts listed are characteristics that support the proposed solution, and tie to a particular standard. These characteristics may be industry standards or industry best practices with respect to the architectural component. Industry standards and industry best practices are tightly related to the business need at hand. That is why it is critical for the agencies to be involved in each of these areas and to understand the requirements.

8.2.1 Harmonization

Harmonizing data is the act of consolidating data from different sources according to the business rules that exist. The illustration below details a simple update where two agencies have mostly different data. Both updates are accepted. An audit of who updated which field is captured to ensure the information is consistent and the updates are traceable.

In this particular situation, the only fields in common are the key to the information, the tax id and company name. In this example, Agency 2 updated the company name. All other fields are not shared between the agency transactions.



Business

Tax ID: 1234
Name: Auriba Corporation
Address Line 1: 890 Hillview Court
Address Line 2: Suite # 130
City: Milpitas
California: CA
Zip Code: 95035

Business

Tax ID: 1234
Name: Auriga Corporation
Category: Software Consulting
Revenue: $$$$$$

Business

Tax ID: 1234
Name: Auriga Corporation
Address Line 1: 890 Hillview Court
Address Line 2: Suite # 130
City: Milpitas
California: CA
Zip Code: 95035
Category: Software Consulting
Revenue: $$$$$$

Audit
Field: Name
Before: Auriba Corporation
After: Auriga Corporation
Updater: Agency 2
Last Updated: 02/31/2009

Figure 8-1 – Data Harmonization

The business rules could have required the common data to be checked to ensure that no updates occurred, it could have ignored updating of the name field or, as in this case it could have taken the update to the name field. The data governance committee will decide which updates are acceptable and which are not. The system should support these complex rules.

Several concepts are discussed in this section. They are:

- data consolidation
- data ownership
- data update rules
- auditing changes
- resolving data update conflicts

The data governance committee decides the rules behind how data is consolidated, owned, updated, and audited, and the rules for resolving the data conflicts that are bound to occur with more complex update rules.

- **Data Consolidation** - The first step is to consolidate the data. Each of the updating systems is evaluated, and a data model produced for a superset of data needed by the updating systems. One of the biggest challenges is data conversion. For example, addresses are traditionally challenging to consolidate since many systems store addresses differently. Converting numbers stored in text fields into numeric data can be challenging. Another challenge is the precision of the data. Data at a higher precision can always be presented at a lower precision, but problems can occur going the other way. Data may not be convertible depending on what was entered and the situation. If so, it may be necessary to keep multiple representations of the data. The Data Governance committee will decide on the best representation for the common data and the rules behind the best consolidation approach.

- **Data Ownership** - Data ownership should be captured for each of the data elements. Since the data is shared, there will be multiple consumers of the data, but only a handful of potential owners. The ownership may change over time. For example the best source for a person's name and address may be from one department initially, however, another department may receive updated information throughout the year. Both must be taken into consideration. The ownership rules on the data may become fairly complex and may change over time. The recommendation is to start simply and evolve from there as needed. Responsibilities of data owners are defined in the State Administrative Manual Chapter 5320.2[xix].

- **Update Rules** - Similar to data ownership, the update rules can be quite complex as well. These rules as well as data ownership will be defined by the business. For example, the tax board might update a person's address once a year (e.g., at tax time), while the Department of Motor Vehicles (DMV) might do many updates per year (with the registration of a car or when a new drivers license is issued). Update rules can be established that will support the normal data lifecycle for a particular piece of data. The rules include not only the order of an update, but also when it is applied and when it is rejected. If the update is rejected it can be queued up for further investigation. This is discussed in more detail in the conflict resolution section.

- **Auditing** -Auditing is a requirement for this type of system, as the update rules can be very complex. The only way to track down an update issue is to keep accurate information about each transaction. Due to the flexible nature of this system, the information that is kept will have to be highly normalized. On every update or new record, information regarding when the update was done and who did it should be captured for each field. The downside of highly normalized data is that it is difficult to review, and therefore a set of screens must be built to enable easy perusal of the audit records.

- **Conflict resolution** - As the update rules become more complex, the probability of having to manage conflicts increases. Conflict resolution is all about managing transactions and or updates that cannot be applied due to some exception to the business rule logic. Since the master data record now does not reflect the latest data out in the enterprise, this transaction cannot be simply discarded, but must, at a minimum, be evaluated and possibly applied to the master data fully or partially. These failed transactions can be queued up and manually reviewed and updated by the data stewards. Roles and responsibilities for governing the data are discussed on Section 5.

The challenges in trying to harmonize data should be noted. Reconciling differences within data across many State departments and across many application and technologies is far from trivial. In fact, this very issue is a common reason why projects like this fail. To reconcile the differences in data and business processes, the business must wholeheartedly participate in data governance.

8.2.2 Availability

As usage of a database increases, so does the significance of an outage. Rarely used databases may tolerate an outage now and then, but frequently used databases supporting critical business processes need to be available whenever a request is made. As more and more 'users[31]' interact with the Shared Data Space, more of the enterprise is impacted by an outage. High availability needs to be designed into any solution at project inception. For this type of solution, to support *Data as a Service* (DaaS), the availability goal should be at least four nines (99.99) of availability.

Special consideration should be given for the data consumers that cannot tolerate any type of outage such as first responders (e.g., law enforcement, fire department, and emergency medical). The availability of the Shared Data Space must exceed the highest availability requirement of any of the data consumers.

Finally, it should be mentioned that High Availability (HA) for a database is not Disaster Recovery (DR), which will be discussed in a different section in this document. Disaster Recovery contrasts to the concept of High Availability in that it focuses on the transition to another database whenever the current database becomes unavailable. HA focuses on keeping the current database functioning properly whenever a failure is encountered. With that said, DR is a contributor to application availability and should be a part of the overall strategy.

Availability can be calculated as:

Availability = ((Time - Outage) / Time) * 100

The following table details out the industry standards for availability.

Availability Percent	Outage Percent	Outage seconds /month	Outage minutes /month	Outage hours /month
99.00%	1.00%	25,920	432.00	7.200
99.50%	0.50%	12,960	216.00	3.600
99.90%	0.10%	2,592	43.20	0.720
99.95%	0.05%	1,296	21.60	0.360
99.99%	0.01%	259	4.32	0.072
99.995%	0.005%	130	2.16	0.036
99.999%	0.001%	26	0.43	0.007

Table 8-1 – Availability Chart

[31] In this case users can mean end users or agency and department applications. It can be any person or process making a request to the database.

So the downtime acceptable for five nines of availability (99.999% uptime) is only 26 seconds per month. For four nines of availability (99.99% uptime), it is less than five minutes per month. With these numbers, it is clear why availability must be designed in to the solution from the beginning.

8.2.2.1 Outage Types

The term "outage" can mean several different things, however, in its simplest form, and for this strategy, it will be defined as the inability to respond to a request that is made by a 'user' in the agreed upon timeframe. Thus, an outage might represent anything from a performance issue to a system failure.

8.2.2.2 High Availability Architecture

In the past, a database had to run within a single machine. The only way to scale the database was to add memory and processors or move the database to a more powerful server.

Later the concept of replication was established where two or more 'master' databases could be established, but that required complex administration as well as sophisticated update rules to handle simultaneous updates to the same record on more than one of the databases.

Now there are clustered databases, where more than one server can support a single database. Several technologies worth considering are:

- Oracle Real Application Clusters
- DB2 High Availability Feature

8.2.2.3 Performance

Performance is not traditionally an item that is discussed within the topic of High Availability. Nevertheless, users perceive poor performance on par with availability, and therefore the topic of performance and scalability should be addressed.

8.2.2.3.1 Scaling For Performance

It is easier and more cost efficient 'scaling up' to address poor performance with using clustered database architecture. In a clustered database environment, servers of similar architecture can participate in the cluster. When demand starts to approach capacity, two options exist for increasing the capacity. One option is to scale the servers vertically by adding more memory or processors. The second option in a clustered environment is to scale horizontally by simply adding another server to the cluster.

It should be noted that there are usually some limitations on the types of servers that can participate in a cluster. For example, Oracle requires the same operating system and system patches be used for each of the servers participating in a Real Application Cluster; however the amount of memory and the number of processors can be different.

8.2.2.3.2 Tuning For Performance

Another way to increase performance on a database environment is to tune the queries that are being executed within that environment. Poorly written queries have much more impact on an environment then one might think. The performance impact can increase exponentially as the number of transactions increase.

Tuning the infrastructure to support an increased demand is another way to get the most out of the database environment. Examples include predefining what servers the queries run on, restructuring how data is laid out within the storage device, or adding indexes. It is all about using the resources within the environment in the most efficient way possible.

8.2.2.4 Downtime

System downtime can be described as any loss of functionality due to a system outage whether it is partial or complete. It is really from the end user perspective. Users who cannot interact with the system care that they can't do their work, not whether a networking, database, or application server problem caused the failure. Returning to our database example, if a server hosting the database crashes in an un-clustered environment, the database becomes unavailable, and the user experiences an outage. In addition some database maintenance requires the database instance to be unavailable to users. All of these contribute to availability issues.

In the following sections the concept of availability is discussed in detail within the context of different types of outages. At a high level there are only two types of outages: planned and unplanned.

8.2.2.4.1 Planned Outage

Most organizations do not count planned outages as a contributor to outage calculations. Businesses with global operations clearly have a need to minimize downtime whether planned or not.

For the State of California and its agencies, a 24x7 operational need may not be a requirement for the Shared Data Space[32]. The State resides in a single time zone, and a maintenance window currently exists for most applications. However, an environment that is highly available during hours of operation is required.

System Changes

Changes to the systems include patching the operating systems on the servers or patching the database software. It also includes hardware changes to the servers, the storage, and the network.

The database architecture identified to support the data sharing strategy should support rolling upgrades and online patching, in addition to being a clustered database environment in the following areas:

[32] Highly available systems exist within the agencies that may meet emergency requirements. Therefore the availability requirement for the Shared Data Space will need to be evaluated based on the agencies and departments involved and the type of data. Since the environment is shared, the most stringent agency availability requirement will need to be adopted for the Shared Data Space.

- Hardware upgrades
- Software upgrades
- Server upgrades
- Database software patching

Currently no database technology supports rolling upgrades and online patching 100% of the time, and that is when the maintenance window is required. However, the technology selected should have robust support for rolling upgrades, and the vendor should be working toward improving this feature.

Data Changes

The database must also support the ability to reorganize the database with minimal downtime. Reorganizations include operations that database administrators do on a regular basis to upkeep the database. The operations include:

1. Rebuilding indexes
2. Dropping partitions
3. Data cleanup
4. Table redefinition

Therefore, careful thought should be given when picking a database technology to ensure that these features are supported.

8.2.2.4.2 Unplanned Downtime

Unplanned downtime is, as the name suggests, unplanned. Something happens to bring the database down, and the goal of the database administrator is to make the application available again to the users as quickly as possible. With new technologies and a clustered database environment, the loss of a server does not necessarily mean a loss of application functionality. It only means a temporary loss in capacity.

Hardware Failures

Hardware failures come in many different forms, however, with the clustered database configurations, redundancy is built into every tier of the configuration. Redundant network cards, redundant networks, redundant servers, redundant connections to storage, and the storage device itself have built-in redundancy. The goal is that any loss of any single component has minimal impact to the end user.

Data Failures

Data failures, like hardware failures, also appear in many different forms. Some can be caused by hardware related issues, software issues, and user related issues. Database technology has matured to address these issues.

Storage Error

Storage area networks are sophisticated storage architecture incorporating multiple storage arrays behind a complex high speed network. The end result is a highly redundant configuration where multiple components must fail simultaneously before an outage is experienced. In addition, the disks are configured in a redundant configuration, the most common configuration for databases being RAID 5 or RAID 10. RAID stands for Redundant Array of Inexpensive Disks (RAID). This section is not intended to be an exhaustive discussion of data storage devices but rather to provide some areas of consideration as the least common denominator for a highly available system[33].

RAID 5 is typically used for data warehouses since its write speed is slower than RAID 10. In the RAID 5 configuration, the disks are configured in a group of typically 1 to 7 disks of equal size and speed. One of the disks has parity information on it so that if any one disk is destroyed the contents of that disk can be rebuilt from the other disks on the fly. In addition, data is stripped across the remaining data disks to minimize ‘hot spots’.

RAID 10 is different in that the arrays of disks are mirrored. Write performance is higher with this configuration, but the overall cost for storage is higher, and therefore it is seen more often implemented for OLTP applications. Like RAID 5, data is stripped across the array and a loss of one disk will not impact the system. A destroyed disk can be rebuilt on the fly from the information on the other disks.

Some databases go a step farther and build some of this redundancy into the database itself. Oracle has a feature called the Automatic Storage Management (ASM) that coordinates the data between groups of disks. This feature reorganizes data on the fly to avoid hot spots and maintains the RAID configuration outside of the storage array.

Human Error

Another factor that should be considered is human error. Even the best database and the best hardware cannot keep user mistakes from occurring. A few human errors to consider are:

- A DBA forgets to use a table in a query
- A user deletes the wrong records from the database and it goes unnoticed for a while
- Poorly written software is used which corrupts or destroys data

Human errors can make the application ‘unavailable’ to the users. Availability, as mentioned earlier, is not really system availability, but the application availability. If the application is unusable due to something that was done by a user, then this is considered an outage. These types of issues are many times not transparent to the user and they are not predictable.

Most of the new database technologies have the ability to rewind the transactions on the database. For example, if data is destroyed by a poorly written software routine, the work that was performed on the database can be rewound to put the data back to a state where it was before the routine was executed. With Oracle technology, these features

[33] RAID configurations come in many different configurations and recently with the improvements with SAN storage the lines between the RAID configurations have blurred. With storage prices dropping it, becomes ever increasing more difficult to justify using RAID5 over RAID 10.

are called Flashback Database, Flashback Table, Flashback Transaction, and Flashback Query.

Data Block and Database Corruption

From time to time, every database administrator is going to experience data corruption, whether it is in the form of a corrupted storage block or a corrupted database. Enterprise Relational Database Management Systems (RDBMS) provide a means of being resilient in case there is a corruption. In the case of corrupted blocks, the database may not be impacted at all, and in the case of a total corruption of the database, a means of point in time recovery is supported by the RDBMS. Each database vendor addresses these issues of block and database corruption in a variety of ways, nevertheless the ability to be resilient to block corruption is a requirement for any highly available system.

8.2.3 Disaster Recovery

Disaster recovery[34] (DR) is the ability to get an application up and running with a complete loss of a datacenter. This type of loss can be in many forms. It can be as simple as losing connectivity to the datacenter due to a connectivity issue (e.g., a backhoe cuts cable) or actually losing a datacenter due (e.g., an earthquake). In either case, to the end user the application is inaccessible.

One of the considerations when designing a DR plan is to consider not only the procedures when the data center is lost but also the procedures when the data center is restored. For example, if connectivity to a data center is lost and all user traffic is now directed to the backup datacenter, how do you reconcile the data now? How long of an outage are you looking at to resynchronize everything? It is the businesses responsibility to determine the DR requirements as well as working with IT to select an approach that meets those requirements.

8.2.3.1 Recovery Terminology

What are RTO and RPO? RTO stands for Recovery Time Objective. This value indicates how quickly recovery has to occur so that the system is available again for use. Long RTO allows for some manual steps to exist in the recovery process. A short RTO is an indicator that the recovery process should be fully automated.

RPO stands for Recovery Point Objective and it is simply ‘How much data are you willing to lose if the system goes down?’ If the answer is a day’s worth of data then the nightly backup will be sufficient for recovery. If the answer is none at all then a synchronous replication solution is required where updates are made to both the primary sites data and the backup sites data. Failure to apply the change on the backup site will reverse the change on the primary site. Most of the time the answer to this question of RPO is to have ‘minimal data loss’, minutes rather than hours or days of transactions. These questions must be answered by the business and not by IT.

As long as some data loss is acceptable then an asynchronous solution can be employed. Synchronous solutions can be problematic in that they require data to be

[34] Also may be known as an operational recovery plan (ORP) or a business resumption plan (BRP).

applied to both databases before the transaction can complete. This can hurt the overall performance.

8.2.3.2 Types of Disaster Recovery

The types of disaster recovery that are employed by companies and the government can be vastly different. There are a host of options out there in the marketplace, each one with benefits and limitations. This section briefly discusses a few of the options. The sections below are not an exhaustive discussion of the entire options available but rather a brief discussion providing just the highlights of the alternatives available. For every example below it is assumed the RPO is not zero.

8.2.3.2.1 Disaster Recovery Hosted Datacenter

In this example, the ‘right’ to a portion of a data center is purchased in case of a disaster. Backup tapes from the night before are restored to similar servers which are leased. To keep cost down these servers are not dedicated to only this customer. Server images along with database backups are restored by administrators to provide a working system. The applications and databases are brought up, briefly tested and user traffic is now directed to these servers.

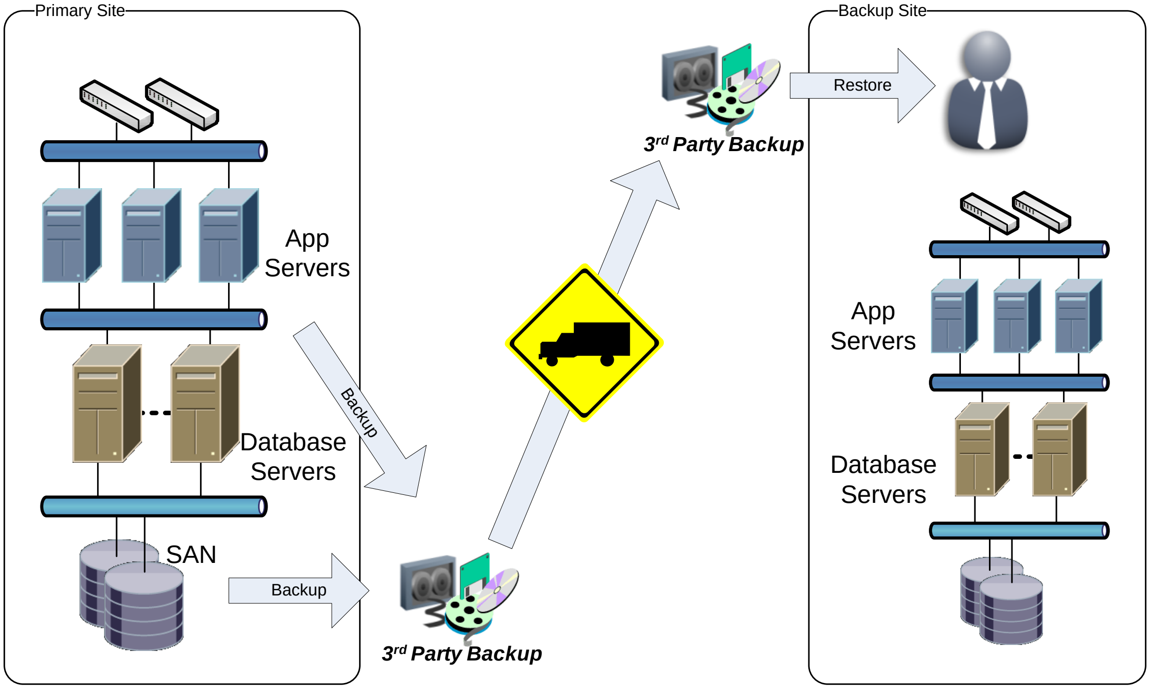


Figure 8-2 – Manual DR Site Illustration

Benefits

1. It is the cheapest solution available.
2. No need for high speed connectivity needed between data centers.

Limitations

1. You are purchasing the 'right' to join the queue of companies to use the servers and databases. The biggest issue is, that since this is not a dedicated resource if a significant disaster occurs (e.g., earthquake) you may be waiting in a queue with other companies for the resources to become available.
2. The other issue is that it does take a considerable amount of downtime before all of the applications that are needed are restored. If the database that is to be restored is quite large, just pulling the backed up database off of tape can take a considerably long time.
3. Reconciling the data after the primary data center comes back on line is also difficult as the entire active user data must be backed up and then restored at the primary site. This in and of itself can cause significant downtime as the application cannot be available during the restore process.
4. Final limitation to consider is that you can only restore since your last backup. Rather than up to the minute image of your user data what you have in this scenario is last night's image of the user data.

8.2.3.2.2 Cold Disaster Recover Site Using Storage Area Network Replication

In this scenario companies have servers available at the backup site but they are turned off. This is normally done to save on software licensing costs. The Storage Area Network (SAN) replicates the data to a similar SAN (e.g., EMC to EMC). If an outage occurs then the backup servers are brought up and user traffic is directed to the backup data center.

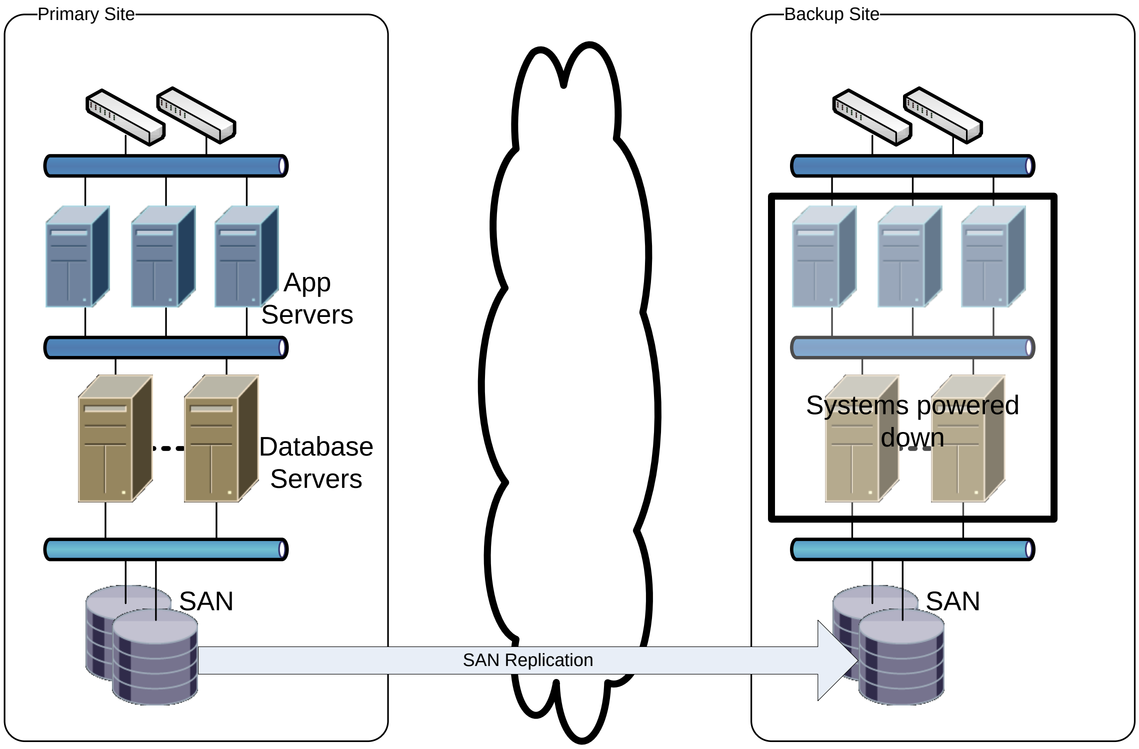


Figure 8-3 – Cold DR Site Illustration

Benefits

1. A cold DR site is a simpler configuration than the hot/active DR site. It is easy to setup and simple to maintain. The SAN does all of the heavy lifting in this configuration.
2. Backups can be made off the replicated copy of the data eliminating traffic on the primary SAN. Although bandwidth for a SAN is rarely a problem the backup can be isolated from the production environment.
3. Rather than last night's data, up to the minute changes are replicated to the other SAN so the users loose less data. In essence only the transactions that were being applied and not committed are lost.
4. In addition, replication technology can introduce a slight delay (e.g., 1 hour) can be configured so that user errors can be recovered from the backup providing they were caught and resolved prior to the mistake being replicated to the other SAN.
5. Some financial savings are realized by not having the servers running in a production environment. Some software vendors will allow their software to be loaded without charging for the installation as long as the application is not ran for some threshold of time.
6. Another benefit is a fairly quick cut over in case of a disaster. All that needs to be done is to let the SAN replication to finish and to bring up all of the servers and redirect the user traffic to the backup systems.

Limitations

1. One major limitation with this approach is that block corruptions are also replicated to the backup SAN. So if there is a problem with the data that was introduced by a hardware or software issue is typically not caught[35]. Many times since this is not a user error the problem is not noticed until after the data is replicated.
2. Another limitation is the hardware sits idle while waiting for a disaster to occur. The servers cannot be easily used for anything else while they are in this configuration.
3. Failover testing becomes a challenge since a contractual time window is in place for the yearly use of the backup servers before a cost is realized for the software.
4. High speed connectivity is needed to properly replicate the data between the SANs.

8.2.3.2.3 Hot Disaster Recovery site using database features

In this scenario the database manages the replication of the data to the disaster set of servers. There are several mechanisms for doing this depending on the technology being used. Most enterprise databases cater for this and allow some level of flexibility to what you can do with the backup servers.

[35] Some block corruption is not apparent until the block is accessed again by the RDBMS system.

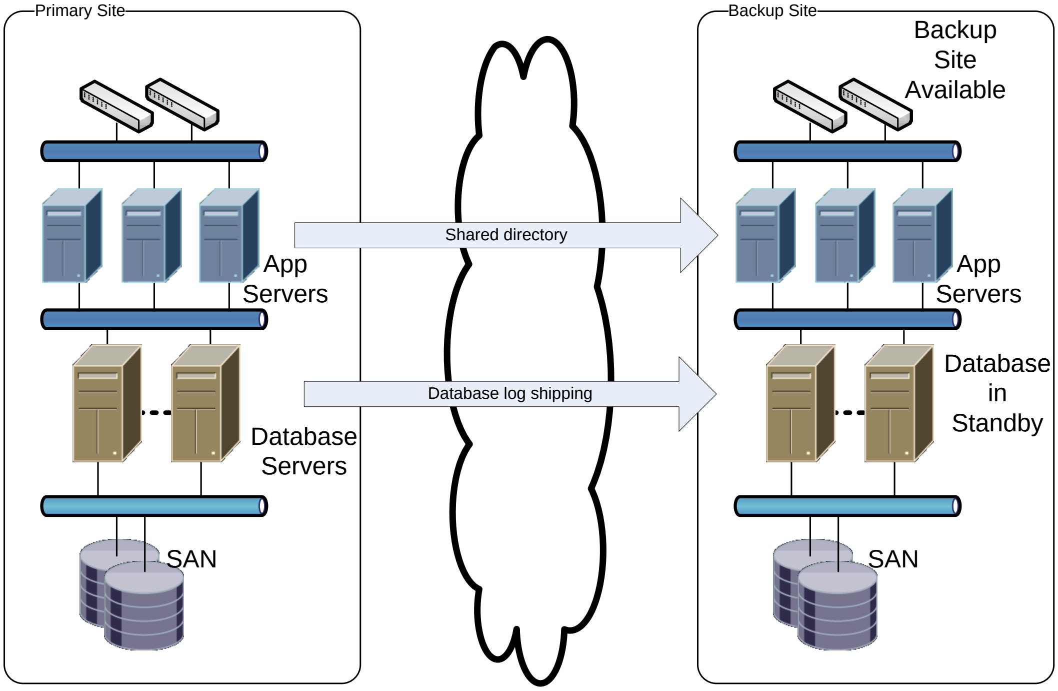


Figure 8-4 – Hot DR Site Illustration

Benefits

1. A hot DR solution has the fastest recovery time over any of the other options discussed. Since the backup set of servers are running, ready to go, once the user data that is 'in flight' is applied then user traffic can be redirected. This scenario brings recovery time to a few minutes and the whole process can be automated.
2. Another benefit is testing out the DR switchover. In this scenario the switchover can be easily tested. Since similar configurations exist in both data centers, the administrators can test the switchover by running for a day on the backup set of servers. Once the day completes then they can simply switch back. This test can be performed on a regular basis and the RDBMS system handles the switch and reestablishes the backup server as the primary and vice a versa.
3. RDBMS vendors have been adding functionality into their products to allow the backup environments to have multiuse. With Oracle 11g the backup database can serve as a read-only repository while applying the log files. This allows reporting traffic to be off loaded to the DR site.
4. Finally the RDBMS system will validate the block changes before they are applied on the backup site. This provides early warning of a block corruption allowing the administrator to intervene before the corruption is applied to the backup database.

Limitations

1. High speed connectivity is needed to properly replicate the data between the database servers.

2. More expensive solution as your backup servers must be active and available. Some vendors do provide a break in the software pricing but nevertheless not only does the hardware need to be purchased but also the application software.

8.2.3.3 Summary

Feature	Hosted DR solution	Cold DR site	Hot DR site
Need for High speed Link	***No***	***Yes***	***Yes***
Expensive	***Low***	***High***	***High***
Block corruption checking	***No***	***No***	***Yes***
Delayed data replication	***Yes***	***Yes***	***Yes***
User data availability	***Last nights***	***Up to date***	***Up to date***
Ease of Testing	***No***	***No***	***Yes***
Risk	***High***	***Low***	***Low***
Multiuse DR servers	***No***	***No***	***Yes***
Switchover Time	***Days/Weeks***	***Hours***	***Minutes***

Table 8-2 – Disaster Recovery Alternative Summary

8.2.3.4 Considerations

The State should consider going with an industry standard enterprise class database like Oracle or DB2 and maintain a hot DR site. Off load any ad-hoc reporting to the DR site to alleviate any impact from the production database.

Another consideration is a modification of the cold standby option where the ***test*** or ***staging*** environment gets re-appropriated as the new production environment in case of a disaster. Switchover time is still in hours however the overall cost is greatly reduced since the environment is multi used and therefore never sits idle. One concern with this approach is the complexity involved with performing an environment switch over. Care to preserve the testing environment must be given while performing a stressful production switch-over.

8.2.4 Information Lifecycle Management

Information Lifecycle Management (ILM) is a storage management strategy which takes advantage of data access requirements becoming less demanding over time (how frequently data is requested and how quickly data needs to be available).

For example after a customer transaction completed, the data will be required for end of month processing (such as billing and tax filing) after which the data is stored for audit and compliance reasons.

The reduction in storage hardware costs through technology and manufacturing advances (e.g., Dollar per GB of storage) is not able to keep up with the growing business demand for the amount of data to be stored.

In the simple example below the amount of data triples roughly every two years resulting in a significant increase of overall storage cost over time (the numbers below are based on $26/GB[36] and are for illustration purposes).

Year	TB	Cost
2009	0.5	$13,000
2011	1.5	$39,000
2013	4.5	$117,000
2015	13.5	$351,000
2017	40.5	$1,053,000

Table 8-3 – Storage Cost Growth

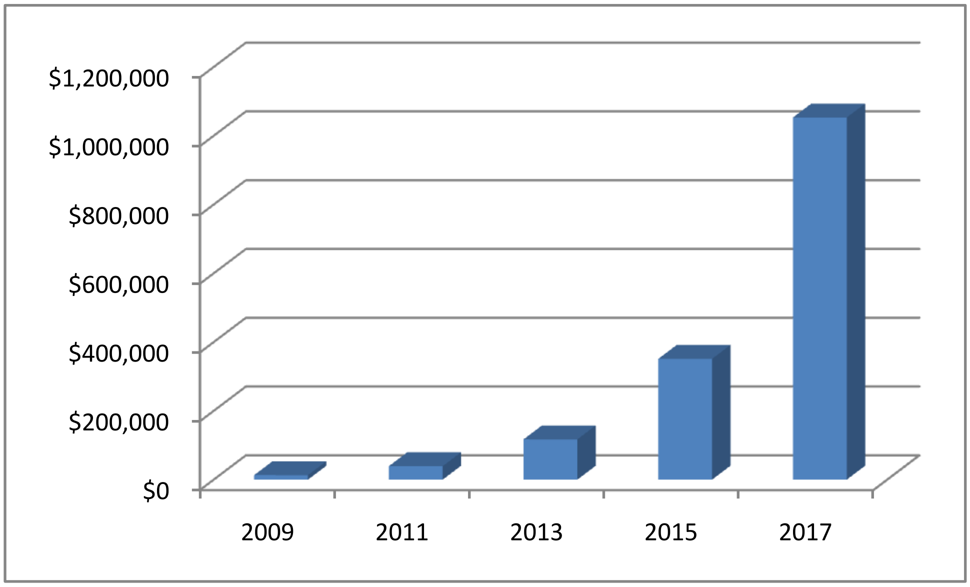


Figure 8-5 – Storage Costs

8.2.4.1 Storage Management Policies

Business drivers determine storage management policies for data criticality (availability and speed of access), confidentiality (sensitivity of the data), recoverability (time to restore), and compliance with internal or external Service Level Agreements (SLA).

Storage management policies in turn drive operational procedures such as data replication, data protection, disaster recovery, and long-term retention as well as infrastructure strategies like storage platforms, network design, and data center strategy (e.g., multiple points of presence, secure building access, and so forth).

[36] $26/GB is OTech cost for their best storage solution.

8.2.4.2 Storage Partitioning

A simple example for a storage management policy is storage partitioning based on data activity using three layers:

1. **High Activity – Class A Storage -** Data is stored in a high speed disk array utilizing fiber optic connectivity to maximize throughput. Only a small amount of the data (~5% e.g., a few days) needs to be accessible in real-time, such as over the Internet, to avoid a potential loss to the business. The data can be partitioned using a date field or a numeric key, with older data being automatically moved to the next storage level, which makes the disk space available for new data.
2. **Low Activity – Class B Storage -** Data is stored on medium speed commodity hardware such as a networked storage appliance. About a third of the data (e.g., last month) needs to be available to the business and customers for accounting and reporting purposes. The majority of the business processes require data in a timely manner but data requests do not need to be immediately served (e.g., batch processing), short delays are acceptable.
3. **Historical – Class C Storage -** Data is stored on low speed disks or might even be farmed out to a tape library. The majority of the data (~60%) has satisfied most business processes and customer needs after which the data needs to be available for auditing and compliance reasons. Internal and external customers are willing to accept that so called old data requests may take a little longer (e.g., data from previous years) as the data has been archived.

Each class of storage in the example above has a different cost factor and the overall amount of data required to be stored differs between the storage classes resulting in a significant storage cost reduction.

Storage Class	Percentage of Data	Dollar per GB	Total Cost
Class A	5%	$26	$17,550
Class B	35%	$12	$56,700
Class C	60%	$8	$64,800

Table 8-4 – Storage Cost by Class

Using the 2015 numbers from Section 8.2.4 the total cost of $351,000 to store 13.5 TB of data can be reduced to about $ $211,950 by applying storage partitioning.

8.2.4.3 Storage Compression

Being able to compress data before it is stored uses less storage and in turn reduces the cost. Modern technologies are capable of compressing data to about a third of its original size, which directly affects the storage cost.

Storage Class	Percentage of Data	Dollar per GB	Total Cost
Class A	5%	$26	$5,850
Class B	35%	$12	$18,900
Class C	60%	$8	$21,600

Table 8-5 – Storage Cost by Class with Compression

Applying both storage partitioning and storage compression techniques to the 2015 numbers from Section 8.2.4 the total cost of $351,000 is reduced to about $46,350, a much more manageable budget.

8.2.4.4 Business Alignment

Information Lifecycle Management differs from traditional Hierarchical Storage Management (HSM) in such that it broadens storage partitioning criteria to elements other than age, such as for example confidentiality and service level agreements, which in turn results in a better alignment with business processes.

The idea is to give the business the flexibility to drive technology decisions instead of having the technology dictate business processes. In a simple example this allows one business division to offer high speed data access to older data if their customers are willing to pay for it, while the remaining divisions run with the regular storage solution.

8.2.4.5 Considerations

The State should consider going with an industry standard database like Oracle or DB2 that will support ILM or to ensure ILM is supported in the storage solution. With the State as large as California and the potential for data growth that exist, the solution to support the data sharing environment must be able to exploit tiered storage solutions.

8.3 Proposed Standards

This section details recommendations where there may be industry standards in place. Whereas database technologies are somewhat proprietary the Open Source and Java communities have been laying out standards in the middleware arena for many years. These standards are evaluated along with standards around Enterprise Content Management, ETL and Trading Partner Agreements.

8.3.1 Security Standards

The State Office of Information Security (OIS)[37] is in the process of identifying and defining standards for security. Since the security standards are under evaluation, a list of security standards considered industry best practices has been provided in the Data Strategy. However, once the set of standards is identified by the State, it will be used in the implementation phase for Data Strategy. OIS has identified that the State agencies must use American National Institute (ANSI) and FIPS standards when sharing data[xx].

Authentication - Authentication is the process of verifying an identity claimed by a system entity. It consists of two steps, identification and verification.
Some of the widely deployed methods for authentication are:

- username / password
- PKI digital signatures (implemented in various technologies - WS-Security, SSL, etc.)
- Kerberos
- SAML
- LDAP
- RADIUS

Authorization - An authorization is a right or a permission that is granted to a system entity for access to a system resource. The common models for designing and implementing security policies are:

- Discretionary Access Control (DAC): where the identity of the requestor is stored together with its permissions
- Mandatory Access Control (MAC): where access is regulated based on a mandated regulation determined by a central authority
- Role-Based Access Control (RBAC): where users are grouped together into roles and permissions are assigned for each role.

Confidentiality - Confidentiality is about information not being made available or disclosed to unauthorized individuals, entities, or processes. It is achieved by means of encryption.

Integrity - Integrity is about ensuring that data has not been changed, destroyed, or lost in an unauthorized or accidental manner. In practice, implementations for determining data integrity rely on hashing algorithms and digital signatures.

[37] OIS was previously a part of Office of Information Security and Privacy Protection (OISPP). OISPP has been split into two separate offices during recent restructuring.

Non-Repudiation - Non-repudiation is the concept of ensuring that a party in a dispute cannot repudiate, or refute the validity of a statement or contract. The most important methods to achieve non-repudiation are: digital signatures, confirmation services and time-stamps.

Privacy - Privacy is the right of individuals to control or influence the collection and storage of information about them. There are no technologies for dealing with privacy.

Availability - Availability is about information being accessible and usable upon demand by an authorized entity.

8.3.2 Enterprise Content Management

Enterprise Content management systems have been evolving for many years. Common features within an Enterprise Content Management offering include:

- Document Management
- Records Management
- Email Management
- Workflow
- Web Content Management
- Archiving
- Digital Asset Management
- Collaboration

A few standards that should be considered when evaluating an Enterprise Content Management Tool are:

US Department of Defense 5015.02-STD Records Management Standard. Many existing records management vendors support this certification. There are standard test cases that an application must pass for this certification. Areas that are tested include:

- Application documentation
- Setup evaluation
- Creation of file plans
- Filing of records
- Searching for records
- Disposition of documents
- System management
- User management

Considering the criticality of the information that is under records management control this standard is a requirement.

JSR 170: Content Repository Java^{TM} technology API. JSR170 is a standard API that is implementation independent that provides a way to access content within a content repository. This API specification allows you to access the functionality of the content repository via a set of services.

WebDAV – "Web-based Distributed Authoring and Versioning". It is a set of HTTP extensions enabling editing and managing of content within a web browser. Most all ECM toolsets have strong support for WebDAV.

8.3.3 Service Oriented Integration

Over the past two decades technology has constantly changed. Just like how "World Wide Web" revolutionized communication and ways of conducting business in the 1990's, SOI has caused a shift in paradigm towards information exchange in the 2000's. As an increasing number of organizations both Corporate and Government, adopt the usage of SOI, industry's best practices have evolved. One set of best practices were compiled and documented by Net-Centric Operations Industry Forum (NCOIF) for Department of Defense in the document "Industry Best Practices in Achieving Service Oriented Architecture (SOA)" (see Appendix K for the document). Industry leaders such as IBM, Microsoft, Booz Allen Hamilton, Oracle, Unisys, etc. have contributed towards the report that is available here. The following is a high-level list of SOI and web services best practices extracted from the document. We recommend reading the original document for more detailed information.

Vision and Leadership

- Evangelize the benefits of net-centricity, SOI, web services, and transformation.
- Think differently.
- Actively manage the cultural, strategic, and tactical issues of a major paradigm shift.
- Proactively address the cross domain and cross business area issues.
- Team with industry, across military services, and across executive agencies.
- Create and document a business case for SOI.

Policy and Security

- Establish technical standards.
- Establish portfolio management policies and policy/information standards and put them in a standards-based registry.
- Establish application interoperability policy.
- Consider how to benefit from both top-down *and* bottom-up leadership.
- Establish governance, security, reuse, compliance, risk management, and versioning policies.

- Employ multiple security approaches.
- Ensure security is “baked into the solution.”
- Address SOI-unique security considerations.
- Plan for disaster recovery, business continuance, and disaster management.

Strategy and Roadmap Development

- Develop, document and publish your SOI strategy.
- Plan for incremental transformation and deployment.
- Align programs/projects to share services.
- Maintain a vision of shared services but move toward it opportunistically and incrementally.
- Design for connections, change, and control.
- Create a common vocabulary.
- Recognize the importance of cross-enterprise architecture.
- Define and enforce application interoperability and business interoperability policies.
- Transform your IT development processes and policies.

Acquisition and Governance

- Incremental acquisition.
- Use experiments, pilots, and collaborative demos.
- Consider using enterprise modeling.
- Enforce policies.
- Loosely coupled services require detailed governance, management, and SLAs.
- Monitor, measure, and analyze the enterprise’s SOI service network.
- Promote Service Discovery and governance using a standards-based registry.
- Consider run-time discovery where appropriate and where it provides business value.
- Promote standards-based process models, such as BPEL or Unified Modeling Language, for process model interoperability.

Implementation and Operations

- Implement incrementally, following the delivery of business value (benefits).
- Partnering and collaborative implementations work best.

- Implementation is more important than theory.
- Pioneer! Do something!
- Ensure a robust publishing and discovery model to facilitate sharing and reuse.

8.3.4 Web Services

Web Services standards have been specified by OASIS (http://www.oasis-open.org), World Wide Web Consortium (http://www.w3.org) and Web Services Interoperability Organization (http://www.ws-i.org). Some of the applicable standards are identified in this sub-section. The following image provides a graphical representation of how all the categories and specifications fit within the context of a Web services framework.

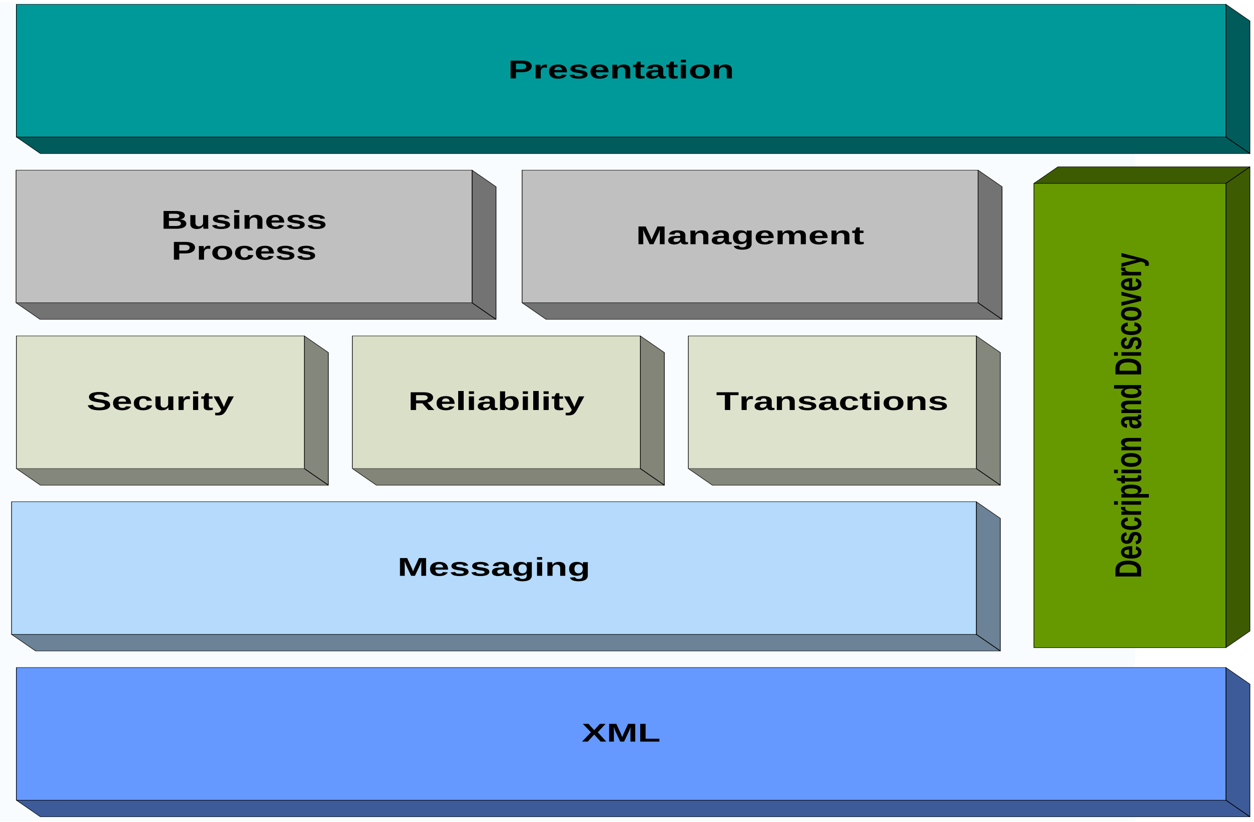


Figure 8-6 – Web Service Standards Stack

8.3.4.1 Extensible Markup Language (XML)

XML is an Open Source standard that is simple and self describing format to encode data or text for systems to exchange and understand.

8.3.4.2 Messaging

The Messaging category focuses on interoperable standards for sending messages between services. These standards could be organized in the following three sub-categories:

a) **Messages:** For services to communicate with each other, messages are encoded according to SOAP specifications, and typically exchanged over HTTP. The SOAP standards are the foundation of network interoperability.

b) **Addressing:** WS-Addressing is a standardized, transport-neutral mechanism that simplifies how two Web Services communicate with each other. It is particularly useful when routing responses to destinations other than the originator.

c) **Notification:** WS-Eventing enables Web services receive messages when events occur in other services and applications. It is a mechanism for a Web Service to register interest (subscription) with another Web service (event source) in receiving messages about events called "notifications" or "event messages".

8.3.4.3 Service Description and Discovery

The infrastructure must enable description of messages and protocols used by Web services using metadata standards. These standards are used by applications and infrastructure to guarantee that services can interoperate based on the requirements services place on users. The important metadata standards include WSDL, WS-Policy, WS-MetadataExchange, and UDDI.

a) **WSDL:** WSDL describes the messages that a service can receive and send. It is the most basic contract language used to describe the business functionality offered by a service.

b) **WS-Policy:** WS-Policy describes the quality of service characteristics and requirements associated with a service.

c) **WS-MetadataExchange:** WS-MetadataExchange is a handshake protocol that allows users to retrieve WSDL and WS-Policy documents associated with a service.

d) **Universal Description Discovery and Integration (UDDI):** UDDI is a model used by service registries. It provides a common repository of metadata about services that can be used to discover what services are available and to select services that are available to use for building new composite services and business processes.

e) **WS-Discovery:** WS-Discovery is a complementary standard to UDDI. It enables run-time discovery of Web Services within a network.

8.3.4.4 Security

WS-Security is a communications protocol providing a means for applying security to Web Services. It enables authentication and authorization by describing how to sign, encrypt and apply security tokens to SOAP messages that ensures end-to-end security. The standards associated to WS-Security are:

a) **WS-Security:** The protocol contains specifications on how integrity and confidentiality can be enforced on Web services messaging. The WSS protocol includes details on the use of SAML and Kerberos, and certificate formats such as X.509.

b) **WS-SecureConversation:** WS-SecureConversation specifies how to manage and authenticate message exchanges between parties including security context exchange and establishing and deriving session keys.

c) **WS-SecurityPolicy:** WS-SecurityPolicy defines how to describe policies related to various features defined in the WS-Security specification.

d) **WS-Federation:** WS- Federation describes how to manage and broker the trust relationships in a heterogeneous federated environment including support for federated identities.

e) **WS-Trust:** WS-Trust describes a framework for trust models that enables Web Services to securely interoperate. It uses WS-Security base mechanisms and defines additional primitives and extensions for security token exchange to enable the issuance and dissemination of credentials within different trust domains.

8.3.4.5 Reliability

How can we ensure completion of message exchanges reliably between participants to solve their business issues? Reliable messaging provides an answer to the question. There are standards in this category that allow messages to be delivered reliably between distributed applications in the presence of software component, system, or network failures. The standard associated is:

WS-ReliableMessaging: WS-ReliableMessaging describes a protocol that allows messages to be delivered reliably between distributed applications in the presence of software component, system, or network failures. The protocol is described in this specification in a transport-independent manner allowing it to be implemented using different network technologies. To support interoperable Web services, a SOAP binding is defined within this specification.

8.3.4.6 Transactions

The Web Services Transactions specifications define mechanisms for transactional interoperability between Web services domains and provide a means to compose transactional qualities of service into Web services applications. It describes an extensible coordination framework and coordination types. The standards associated are:

a) **WS-Coordination:** WS-Coordination describes an extensible framework for providing protocols that coordinate the actions of distributed applications. Such coordination protocols are used to support a number of applications, including those that need to reach consistent agreement on the outcome of distributed activities. The framework defined in this specification enables an application service to create a context needed to propagate an activity to other services and to register for coordination protocols. The framework enables existing transaction processing, workflow, and other systems for coordination to hide their proprietary protocols and to operate in a heterogeneous environment. Additionally this specification describes a definition of the structure of context and the requirements for propagating context between cooperating services.

b) **WS-AtomicTransaction:** WS-AtomicTransaction provides the definition of the atomic transaction coordination type that is to be used with the extensible coordination framework described in the WS-Coordination specification. The specification defines three specific agreement coordination protocols for the atomic transaction coordination type: completion, volatile two-phase commit, and durable two-phase commit. Developers can use any or all of these protocols when building applications that require consistent agreement on the outcome of short-lived distributed activities that have the all-or-nothing property.

c) **WS-BusinessActivity:** WS-BusinessActivity provides the definition of the business activity coordination type that is to be used with the extensible coordination framework described in the WS-Coordination specification. The specification defines two specific agreement coordination protocols for the business activity coordination type: BusinessAgreementWithParticipantCompletion and BusinessAgreementWithCoordinatorCompletion. Developers can use any or all of these protocols when building applications that require consistent agreement on the outcome of long-running distributed activities.

8.3.4.7 Business Process

A business process specifies the potential execution order of operations from a collection of Web services, the data shared between these Web services, which partners are involved and how they are involved in the business process, joint exception handling for collections of Web services, and other issues involving how multiple services and organizations participate. Business Process Execution Language (BPEL) for Web Services specifies business processes and how they relate to Web services.

8.3.4.8 Management

WS-Management describes a general SOAP-based protocol for managing systems such as PCs, servers, devices, Web services, other applications, and other manageable entities.

8.3.4.9 Presentation

Web Services for Remote Portlets (WSRP) is a specification which defines how to leverage SOAP-based Web services that generate mark-up fragments within a portal application. By defining a set of common interfaces, WSRP allows portals to display remotely-running portlets inside their pages without requiring any additional programming by the portal developers. To the end-user, it appears that the portlet is running locally within their portal, but in reality the portlet resides in a remotely-running portlet container, and interaction occurs through the exchange of SOAP messages. Leveraging WSRP within a Service-Oriented Architecture provides a powerful combination whereby presentation-oriented portlet applications can be discovered and reused without engaging in additional development or deployment activities.

8.3.5 Geospatial Web Services

The standards body, Open GIS Consortium Inc. (OGC) is an international non-profit organization that develops Geospatial standards. Some of those standards cover GIS web services. There are two specifications that are referenced in this document are:

- OpenGIS® Web Map Server Implementation Specification[xxi]
- Web Feature Service Implementation Specification[xxii]

The web services identified in the data strategy leverage these specifications. OGC's website can be found at http://www.opengeospatial.org/.

8.3.6 Bulk Data Loads through Extraction, Transformation and Loads

A more efficient means for transferring large amounts of data can be performed through a methodology commonly referred to as Extract, Transform, and Load (ETL). Many toolsets have been produced that utilize this methodology and they are in common use within data warehousing phases. The steps that make up the ETL methodology are:

- **Extraction -** The first step of moving information between systems or making information available to applications such as analytics and reporting. Tools need to be able to read data from a multitude of sources and formats as well parse this data to determine if expected patterns or structures are met.
- **Transformation -** Rules and functions are applied to modify data patterns or structures without loss of information. The modifications may range from simple sorting and filtering, across more complex merging or splitting of data components, to translating keys or codes between systems.
- **Loading -** Previously prepared data is inserted into the destination systems. Load exceptions may occur as the destination system reacts to the inserted data through key or unique constraints and custom defined triggers and stored procedures. The most commonly load exceptions are duplicates and orphans (linked information missing its predecessor).

8.3.6.1 Performance Considerations

A main contributor to successful information sharing is the speed at which ETL can be performed. For example processing 10 million rows at 200 rows per second can be accomplished in under 14 hours. However if a system processes at a slower pace, the ETL will not be able to complete before next day's ETL is supposed to start again.

The slowest part of the ETL process is typically the load operation. To increase loading speed the best practice is to ensure the data is correct before it is loaded and to disable all constraints and scripts while the load is in progress.

Successful ETL requires close participation of all parties involved and is not a standalone effort. Below are some of the performance boosting techniques commonly used:

- Generate keys and code transformations within the ETL tool and outside the target system.
- Leverage partitioning (tables and indexes) to reduce processing time to the range of information to be shared (e.g., for the current day).

- Some target systems have business logic setup which kicks off for every record modified. Such business logic is typically designed for transactional processing (a few operations per hour) and is not tuned for bulk processing (a million rows per hour). To save time simulate the effect of the business logic as a separate step.
- Relational databases leverage data access structures (so called indexes) to quickly find records in large tables. Indexes are similar to a binary tree and need to be rebalanced (to guarantee shortest access paths) when the underlying data is changed significantly. To avoid time consuming rebalancing during ETL operations, indexes of affected tables are removed prior to the load and recreated afterwards. This may result in slow or no system response during the load.
- Many database systems feature parallel operational support, which greatly speeds up ETL. However the nature of parallelism frequently requires single user access to the entire database system (or at least the affected data objects) to avoid conflicts with other data operations. As a result less ETL time is traded for a scheduled maintenance window, meaning a planned down time of the system.
- Perform all data validation (patterns and structures) before the load starts. Disable all data integrity rules on the system into which data is loaded.

8.3.6.2 Spatial ETL

Software capable of performing Extract, Transform, and Load (ETL) operations on geographic or locations data are commonly referred to as Spatial ETL tools. The challenge for extracting and loading spatial data is that it may be stored in a variety of different formats which the software needs to be able to recognize and process correctly.

These challenges are particularly apparent when transforming spatial data between geographies as data attributes vary between resolution levels, the following capabilities are required:

- **Reprojection -** Converting spatial data between coordinate systems
- **Spatial transformations -** Modeling of spatial interactions and calculating spatial predicates
- **Topological transformations -** Creating of topological relationships between datasets
- **Resymbolisation -** Changing of the cartographic characteristics e.g., color or style of lines
- **Geocoding -** Converting of tabular data into spatial data

8.3.6.3 ETL Tool Considerations

There are several advantages with using an ETL tool over conventional means of bulk data transfer. These advantages include maintainability and connectivity. ETL tools provide a consistent development environment regardless of the data source. Because the development environment is consistent, similar ‘code’ can be written that will transfer data from a wide variety of sources. This flexibility does come with a price and there are several items that must be considered when choosing an ETL toolset. The considerations include:

- The Data Access
- The Access Performance
- The Development Environment

8.3.6.3.1 Data Access

At the very minimum an ETL tool taken under consideration should be able to "read from" and "write to" all required data sources (such as relational databases or flat files).

Configuring a data source within the ETL tool should be a simple, straight forward, almost intuitive process. State-of-the-art ETL tools undergo a lot of in-house testing at the vendor side and have their basic workflows streamlined as a result (to save time during quality assurance).

Look for so called "native" support when setting up data sources. Leveraging third party drivers (such as ODBC) saves the ETL tool vendor development time but results in slower data throughput and creates a potential support issue with the driver vendor.

8.3.6.3.2 Access Performance

Being able to access data is only part of the journey, the ETL tool will need to read and write data fast enough to meet SLAs such as daily data loads have 24 hours to complete.

The State should perform benchmark testing during a trial evaluation of the ETL software; this will provide a close to reality impression of the capabilities of the tool.

The hardware setup can have a profound effect on the benchmark tests, simple changes to the OS or introducing RAID may produce vastly different results. It is desirable to use a standalone system for testing only to be able to compare the results for different ETL tools.

8.3.6.3.3 Development Environment

A commonly overlooked aspect of ETL software is, that it provides a development environment which needs to support all required transformation logic. A good ETL tool provides a development environment which helps to reduce the need of knowledge transfer and lowers support costs.

The transformation capabilities may impact performance as well and need to be benchmarked independently. Performance impairment occurs when the code execution is slower than the data access (because complex libraries are used to perform simple operations).

Common features an ETL tool should have:

- Separation of data access from data management (transformation logic should not be bound to the data access code)
- Perform date format conversions natively (not using string operations)
- Fast string operations (obtaining a substring does not rely on large string libraries being loaded into memory)
- Functions for string pattern validation (the ability to verify the format of a field e.g., SSN, phone number, zip code, etc.)
- Library caching (code libraries are not freed from memory after each use)
- Built-in key management and caching (key conversions are supported by the tool and don't require development effort)
- Lookup caching (keep small lookup tables in memory to speed up conversions)
- Support SQL query hints (the ability to let a database know which access path to choose e.g., specific index)
- Support the execution of external programs (such as calling shell scripts and being able to obtain return values from them)
- Support parallelism (leverage multiple cores instead of being limited to single threaded execution)
- Version control support (multiple developers work together; the ability to roll back changes to the last known working version)

8.3.6.4 Success Factors

The success of an ETL tool implementation depends on several characteristics, which every organization weighs according to internal drivers:

- **Speed to market -** How quickly can new solutions or enhancements to existing solutions be rolled out?
- **Functional capabilities -** Processing throughput and processing flexibility are the drivers of the technical capabilities of the tool.
- **Maintenance complexity -** The ability of existing and new staff to interpret and maintain existing transformation logic. Substandard ETL tools necessitate custom code development which may result in a large and often poorly documented code base.
- **Financial ROI -** Implementations of ETL tools must produce a positive return on investment (ROI).

8.3.6.5 ETL Standards

ETL has long been a proprietary solution for transferring data with very expensive tool sets being offered by some of the largest software vendors. Many of these vendors have proprietary approaches to data management. With the emergence of the Open Source community, Open Source ETL frameworks are becoming more prevalent. Two of these Open Source frameworks are:

MIKE2.0 (Method for an Integrated Knowledge Environment) - http://mike2.openmethodology.org
An Open Source framework for enterprise information management designed for: data quality improvement, data integration, data migration, data warehousing, and master data management.

With over 600 articles, MIKE2.0 is a comprehensive framework which provides detailed step by step instructions (300+ steps) to help you with its implementation and roll out: http://mike2.openmethodology.org/wiki/Overall_Task_List

MIKE2.0 Core Solution Offering caters to these Offering Groups:

- Business Intelligence and Performance Management
- Information Asset Management
- Access, Search and Content Delivery
- Enterprise Data Management
- Enterprise Content Management
- Information Strategy, Architecture and Governance

MIKE2.0 Methodology leverages a five phase Continuous Implementation approach for implementation and rollout (instead of the traditional linear or waterfall approach):

- Phase 1: Business Assessment and Strategy Definition Blueprint
- Phase 2: Technology Assessment and Selection Blueprint
- Phase 3: Information Management Roadmap and Foundation Activities, Design Increment, Incremental Development, Testing, Deployment and Improvement
- Phase 4: Repetition of Phase 3 with operational feedback into Information Management Roadmap and Foundation Activities
- Phase 5: Repetition of Phase 3 with operational feedback into Information Management Roadmap and Foundation Activities

Clover.ETL - http://www.cloveretl.org
A Java based Open Source ETL framework with a commercial product line offered by OpenSys (http://www.opensys.com). The Open Source framework allows contributions from a large developer community while the commercial aspect provides support and high quality software.

The Open Source Clover.ETL consists of three main components:

- **CloverServer -** High speed, high quality extension of CloverETL
- **CloverETL -** The ETL core software component
- **CloverGUI -** A graphical user interface (GUI) to visually create and modify data transformations

8.3.7 Standards for Trading Partner Agreements

Several industry sponsored open standards exists for defining trading partner agreements, in the following we outline the two most widely accepted ones.

ISO 15000-1: ebXML Collaborative Partner Profile Agreement - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39972
Part one of ISO/TS 15000 outlines the Collaboration-protocol profile and agreement specification (ebCPP) as part of the Electronic business eXtensible Markup Language (ebXML) framework.

Each trading partner sets up their own Collaboration Protocol Profile (CPP) XML document which describes their abilities in a partner exchange, such as:
Supported messaging protocols
Supported security capabilities

The Collaborative Partner Agreement (CPA) document describes the relationship between two trading partners, such as:

- Identification
- Communication (protocols)
- Service (URLs)
- Security
- Exception handling (e.g., duplicate messages)
- Acknowledgment of receipt

Trading Partner Agreement Markup Language (tpaML) - http://xml.coverpages.org/tpa.html
This is an early effort from IBM of a formal specification for a Trading Partner Agreement Markup Language utilizing XML to implement electronic contracts.

The foundation of tpaML is the Trading Partner Agreement (TPA), which defines how trading partners will interact at the transport, document exchange and business protocol layers.

A TPA contains the general contract terms and conditions, participant roles (buyers, sellers), communication and security protocols and business processes, (valid actions, sequencing rules, etc.).

The tpaML specification covers the following:

- Identification
- Communication (protocols and electronic addresses)
- Security (certificates used for authentication, nonrepudiation, and digital envelope, and other security parameters)
- Non-technical aspects (e.g., the valid duration of the TPA)

- Data Definition (formats)
- Role & Responsibilities
- Action List (definition of message flows between the invoker and the service provider, responsiveness, failure handling, and other attributes)
- Sequencing Rules (describe valid action invocation sequences)
- Global Properties (defaults values)
- Comments (intended for handling of disputes, termination of the TPA, and other exceptional conditions)

8.4 Summary

The solution towards implementing the design and architecture specified in Section 3 requires integration of multiple components. In Section 8 we have covered key both concepts and standards that support the strategy. The concepts covered were those of Harmonization, Availability, Disaster Recovery and Information Lifecycle Management for database management systems. As depicted from the information given, the concepts represent industry best practices that are supported in a similar architecture proposed in the data strategy architecture.

The standards outlined in section 8 of the report are, Security, Enterprise Content Management, SOI, Web Services, Trading Partners and ETL. Although there is a significant amount of material and information on these standards, we felt that these six standards encapsulated the view supported by the strategy.

Appendix A - Data Strategy Implementation Work Breakdown Structure

The purpose of the work breakdown structure is to provide input on the steps and resources necessary for the implementation of each of the phases listed in the data strategy. It should be noted that due to the nature of the work, maturing technology, the evolving business and the vendors that are chosen in the end, this work breakdown structure will need to be revisited. This section provides a template as a start toward a project plan for the phase.

Roadmap Data Strategy Overarchinc

Data Strategy Infrastructure Config

Security Architecture Design Roadmap.pdf

Master Data Repository Roadmap.

Metadata Repository Roadmap.pdf

Content Management System

Enterprise Service Bus Roadmap.pdf

Web Services Roadmap.pdf

Data Warehousing Roadmap.pdf

Trading Partner Network Roadmap.pc

This page intentionally left blank

Appendix B - Industry Standards

National Information Exchange Model (NIEM) – http://www.niem.gov

The National Information Exchange Model (NIEM) is a set of standards and processes leveraging eXtensible Markup Language (XML) for information exchange. NIEM is designed to develop, disseminate, and support information sharing for justice, emergency and disaster management, public safety, and homeland security agencies. The standard is highly extensible and is easily adopted across a wider scope of agencies.

Below is a quick overview of the NIEM features and how they work:

- **Data Components**
  The XML representation of business objects containing all information exchanged between agencies.
- **Information Exchange Package Documentation**
  Data components exchanged between agencies are organized into Information Exchange Packages (IEPs).
  Metadata (e.g., information about the structure, content, confidentiality, and criticality) for a specific type of information exchange between two agencies is captured by the Information Exchange Package Documentation (IEPD).
- **NIEM Core**
  The core components of the XML framework provide specifications for the most commonly used data components (e.g., person, address, and organization), which are reusable across multiple agencies and different subject areas. The NIEM Core is relatively small and doesn't undergo frequent changes.
- **Domains**
  Subject areas with similar applications (and data components) are grouped into so called domains (e.g., health care, supply chain management, etc.)
  Subject Matter Experts (SMEs) act as data stewards for these domains to answer questions about data components and their use.
- **Communities of Interest (COI's)**
  User groups collaborating in their domain. Common activities are to expand existing and create new data components and to establish and refine a controlled vocabulary.
- **NIEM Conformance**
  The NIEM standard provides a set of conformance rules which agencies leverage to setup new and manage existing information exchanges.

The NIEM standard provides the following reference schemas, which are XML schema definitions:

- **NIEM Reference Schemas**
  All schemas with content created and/or approved by the NIEM steering committees which are periodically released in schema distributions.
- **Subset Schema**
  Definitions of subsections of the NIEM Reference Schemas required for a specific information exchange.

- **Support Schemas**
  Helper schemas needed to build NIEM conformant schemas. The three Support Schemas are: APPINFO, STRUCTURES, and PROXY.
- **Extension Schema**
  Domain extension of the NIEM Reference Schemas.
- **Exchange Schema**
  XML schemas for Information Exchange Packages (IEPs) and Information Exchange Package Documentation (IEPD).
- **Constraint Schema**
  XML schemas which provide additional constraints for NIEM-conformant instances.
- **Codelist Schemas**
  An XML schema defining a range of acceptable values for elements of data components.

NIEM supplies free tools which fully support all structural and content features and also enable third-party vendors to develop additional tools:

- **NIEM XML Data Dictionary Spreadsheet**
  http://www.niem.gov/topicIndex.php?topic=spreadsheet

  A Microsoft Excel spreadsheet with all XML schema information for the NIEM Core. The spreadsheet provides a comprehensive description of each of the NIEM XML nodes.
- **Schema Subset Generation Tool**
  http://niem.gtri.gatech.edu/niemtools/ssgt/index.iepd

  A tool to assist in building NIEM Subset schemas.
- **Information Exchanges Mapping Tool**
  http://niem.gtri.gatech.edu/niemtools/mapping/index.iepd

  A tool for defining Information Exchange Package Documentations (IEPDs).
- **IEPD Tool**
  http://niem.gtri.gatech.edu/niemtools/iepdt/index.iepd

  A tool to enter and upload metadata for Information Exchange Package Documentations (IEPDs).

OpenDocument Format (ODF) – http://www.openoffice.org

The OpenDocument Format (ODF) is a set of XML schema definitions to store commonly used office documents such as word processing, spreadsheets, presentations, and databases.

Binary data (e.g., pictures, data, etc.) is located in a subdirectory under the main XML document which is stored in a single compressed archive file. All main OpenDocument files have as their root element the `<document>` XML tag.

ODF is a free and open standard developed by the Organization for the Advancement of Structured Information Standards (OASIS), and is also published as an ISO/IEC international standard (under ISO/IEC 26300:2006 Open Document Format for Office Applications v1.0).

The most commonly known office suites supporting the ODF standard are:

- AbiWord
- Corel WordPerfect Office X4
- Google Docs
- IBM Lotus Symphony
- KOffice
- NeoOffice
- OpenOffice.org
- SoftMaker Office
- Star Office
- Zoho

SOAP – http://www.w3.org/TR/SOAP/

Originally referred to as Simple Object Access Protocol (SOAP) but now it is referred to only as SOAP. It provides the foundation of a messaging framework utilizing an XML protocol standard to exchange information for Web Services.

The World Wide Web Consortium (W3C) defines a Web Service as “a software system designed to support interoperable machine-to-machine interaction over a network”. Below is a simple Web Services example utilizing the SOAP protocol:

1. A computer system sends an automated SOAP message to a web service enabled web site. The SOAP message includes data required for processing (such as a catalog lookup request).
2. The web site responds with a SOAP message (an XML formatted document including the resulting data such as prices, location, features, etc.)
3. The returned data can be automatically processed by the initiating computer system without further human intervention.

SOAP leverages other application layer protocols like Simple Mail Transfer Protocol (SMTP), Hypertext Transfer Protocol (HTTP), or Remote Procedure Call (RPC) when transmitting the SOAP message. The SOAP specification is currently maintained by the XML Protocol Working Group of the World Wide Web Consortium (W3C).

The latest version of SOAP is version 1.2 and consists of three main parts:

1. The SOAP 1.2 Primer (which is part zero of the SOAP specification) is an introduction to SOAP providing an easy to understand tutorial.

2. Part 1 of the 1.2 specification defines the SOAP messaging framework consisting of:
   - **SOAP Processing Model**
     Defines the rules for processing a SOAP message
   - **SOAP Extensibility Model**
     Defines the concepts of SOAP features and SOAP modules
   - **SOAP Protocol Binding Framework**
     Describes the rules for defining a binding to an underlying protocol used for exchanging SOAP messages between SOAP nodes
   - **SOAP Message Construct**
     Defines the structure of a SOAP message
3. Part 2 includes appendices for the SOAP messaging framework.

Advantages of SOAP are:

- SOAP is platform independent (e.g., Intel, Unix, etc.)
- SOAP is language independent (e.g., Java, C#, C++, etc.)
- SOAP is simple and extensible (custom XML tags such as confidentiality and criticality can be added as needed)
- SOAP can be over a multitude of transfer protocols. For example SOAP over HTTP allows access through network proxies and firewalls (and doesn't require additional ports to be opened by your networking team)

WSDL – http://www.w3.org/TR/wsdl

The Web Services Description Language (WSDL) is a language utilizing XML documents to describe web services and how they are accessed by web based applications. WSDL can be thought of as an electronic catalog of available SOAP services.

WSDL XML documents define services as sets of network endpoints, or so called ports. A port is defined by associating a network address with a reusable binding. A set of port definitions specifies a service. Messages are described by the data they may exchange, while port types are groups of supported operations. The definition of ports and their messages are reusable for similar web services.

For example:

1. A computer system connecting to a web service enabled web site utilizes WSDL to first obtain a catalog of available services.
2. The web site responds with a WSDL XML response file containing all necessary information to access any of the available web services.
3. The initiating computer system can then use SOAP to further interact with web services listed in the WSDL. All this is accomplished automatically without human intervention.

WSDL was initially developed by Microsoft, Ariba, and IBM, who submitted version 1.1 of the specification to the World Wide Web Consortium (W3C).Initially the W3C accepted WSDL only as a note but published the standard on their web site. Twenty-two other companies joined the submission, which at that time was the largest number of W3C members to support a joint submission. WSDL therefore already enjoys broad-based support, and there are now many other vendors who provide implementations of WSDL in their Web services products.

The latest version is WSDL Version 2.0; its specification consists of three parts (following the SOAP Version 1.2 outline):

1. The WSDL 2.0 Primer (which is part zero of the WSDL specification) is an introduction to WSDL providing an easy to understand tutorial.
2. Part 1: Core of the 2.0 specification defines the WSDL framework.
3. Part 2: Includes appendices for the WSDL framework.

WS-Security – http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wss

http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-SOAP-message-security-1.0.pdf

WS-Security (Web Services Security) is an expansion of the SOAP messaging framework to include security features in the header of SOAP messages. End-to-end security is ensured by allowing only the application layer to access the security content.

WS-Security describes how to attach signatures and encryption headers to SOAP messages. In addition, it describes how to attach security tokens (such as X.509 certificates or Kerberos tickets) to messages.

WS-Security was originally developed by IBM, Microsoft, and VeriSign. Since then Oasis-Open has taken over development. The latest release was version 1.1 in February of 2006.

WS-Reliability – http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrm

http://docs.oasis-open.org/wsrm/ws-reliability/v1.1/wsrm-ws_reliability-1.1-spec-os.pdf

WS-Reliability is an expansion of the SOAP messaging framework to include reliable messaging requirements in the header of SOAP messages. This expansion is leveraged when an application must also guarantee reliability and security when exchanging information between web services.

This specification has been designed for use in combination with other protocols and relies on additional services such as the ebXML Message Service.

WS-Transaction – http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=ws-tx

http://www.oracle.com/technology/pub/articles/dev2arch/2004/01/ws-transaction.html

WS-Transaction is an expansion of the SOAP messaging framework to describe coordination types in the header of SOAP messages, which are used with the extensible coordination framework described in the WS-Coordination specification.

WS-Transaction specifies two coordination types:

- Atomic Transaction (AT) for individual operations, and
- Business Activity (BA) for long running transactions.

Developers can choose either one or both of these coordination types when building applications that require data consistency across distributed transactions.

The latest version of the standard (WS-TX 1.1) was released by the Organization for the Advancement of Structured Information Standards (OASIS) in July 2007 and includes the following parts:

- WS-Coordination
- WS-AtomicTransaction
- WS-BusinessActivity

eXtensible rights Markup Language (XrML) – http://www.xrml.org

The eXtensible Rights Markup Language (XrML) is a language to manage digital rights. It utilizes XML documents to describe rights, fees, and conditions as well as message integrity and entity authentication data.

The following is a list of advantages of using XrML:

- XrML is independent of media type, format, platform, and delivery mechanism.
- Provides a high level of security by digitally signing all XrML labels and licenses.
- Allows to define entities which provides interoperability across multiple platforms and applications.
- Leverages other standards to specify digital signatures, digital identifiers, content metadata and so forth.
- The XrML framework comes with tools, tutorials, documentation, examples, and use cases.

XrML is supported by the MPEG-21 and OASIS Rights Language Technical Committee (RLTC). This includes direct support from companies such as Cisco, HP, IBM, Universal Music Group, and VeriSign. There are also several products already using or planning to use XrML from companies such as Contents Works, DMDsecure, Integrated Management Concepts, Microsoft, OverDrive, Sony, and Zinio.

Universal Description Discovery and Integration (UDDI) – http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=uddi-spec

http://www.oasis-open.org/committees/uddi-spec/doc/tcspecs.htm

The Universal Description, Discovery and Integration (UDDI) is an open industry initiative to support a worldwide registry for businesses to publish their electronic service listings and to define how automated computer systems may interact with each other over the Internet.

UDDI responds to Simple Object Access Protocol (SOAP) message requests with Web Services Description Language (WSDL) documents describing the protocol bindings and message formats required to interact with the web services listed in its directory.

Servers which support the UDDI specification and belong to a UDDI registry are called UDDI nodes. Some UDDI nodes may also serve as UDDI registries.

A UDDI business registration consists of three components:

- **White Pages -** Provide data such as contact, address, and known identifiers
- **Yellow Pages -** Provide classification by industry based on standard taxonomies
- **Green Pages -** Provide technical information about offered web services

ebXML – http://www.ebxml.org

Electronic Business using eXtensible Markup Language (ebXML) is set of standards and processes outlining an electronic business framework with the goal to enable a global electronic marketplace where enterprises regardless of size and geographic location may do business with another by exchanging XML messages.

The ebXML framework is capable of dealing with various reliability, legal, and international issues when exchanging business documents. ebXML is a joint initiative of the United Nations (UN/CEFACT) and the Organization for the Advancement of Structured Information Standards (OASIS).

Some of the advantages of using ebXML are:

- Comes with plug and play style shrink-wrapped solutions
- Enables all parties (regardless of size) to engage in Internet-based electronic business
- Enables parties to complement and extend current EC/EDI investment to expand electronic business to new and existing trading partners

The International Organization for Standardization (ISO) has approved the following five ebXML specifications as the ISO 15000 standard:

- ISO 15000-1: ebXML Collaborative Partner Profile Agreement
- ISO 15000-2: ebXML Messaging Service Specification
- ISO 15000-3: ebXML Registry Information Model
- ISO 15000-4: ebXML Registry Services Specification
- ISO 15000-5: ebXML Core Components Technical Specification, Version 2.01

ISO/IEC 11179 – http://metadata-standards.org/11179

ISO/IEC 11179 describes how organizations can standardize and register metadata elements within a worldwide metadata registry to make their information understandable and shareable. The goal is to enable a worldwide open exchange of data by electronic information interchanges.

Metadata about data elements is stored in a data element registry. A data element registry supports information sharing with descriptions of data. Registration is the process of documenting metadata to support sharing of information. Registration is carried out at the data element level which maximizes its semantic value. ISO/IEC 11179 enables users to unambiguously interpret the intended meaning of information.

The ISO/IEC 11179 standard supports the following:

- Electronic information sharing within an agency and across different agencies
- Acquisition and registration of data
- Simplifies data manipulation by leveraging its metadata
- Bridges software, hardware, geographic, and organizational boundaries

The ISO/IEC 11179 standard consists of six parts:

1. Framework
2. Classification
3. Registry metamodel and basic attributes
4. Formulation of data definitions
5. Naming and identification principles
6. Registration

eXtended MetaData Registry (XMDR) – http://xmdr.org

https://xmdr.lbl.gov/mediawiki/index.php/Main_Page

The Extended Metadata Registry (XMDR) is an initiative to extend the ISO/IEC 11179 standard with the goal to improve storing and retrieving of semantics for data elements, terminologies, and concept structures.

Some of the goals of the XMDR initiative are:

- Improve representation of relationships between data
- Register and manage complex semantic metadata
- Add more rigorous and formal specifications
- Use concepts to unify different types of metadata

Metadata Registry

A growing emphasis on standardizing the shareable information assets within an organization, and information interchange, both within and with external sources, is driving organizations to focus on their information architecture strategy.
Considering the wide spectrum of shareable information, the right information architecture is to setup a central repository or a “virtual” central portal that accounts for definitions, values (in some cases) and the change management processes around the quality of information being maintained. An alternative to creating a central repository is to create a central metadata registry – which provides an organization a more structured framework to work within, and also addresses the diversity inherent in shareable data assets.
Standards for models and templates for metadata registries already exist – for example, the ISO 11179 standard for Metadata Registries, Dublin Core, and ebXML for XML registries being a few. Depending on the complexity of the model, implementations of registries can be quite challenging.

What is a Metadata Registry

A Metadata Registry is defined as an automated resource “used to describe, document, protect, control and access informational representations of an enterprise”. There are various interpretations of what is held in a metadata registry:

- Standardized information in a pre-defined model
- Metadata, system metadata, system engineering
- Reference information

A typical metadata registry has the following characteristics:

- A generic model to store all the information
- A ‘formal’ registration process that allows elements or objects to be properly ‘registered’ (i.e., accepted) in the registry
- Layered organization and responsibility structure for approvals and standardization
- Strong stewardship and security controls
- Different options to present the information to the users in order to facilitate the main objective of the registry i.e., information sharing.
- Extranet access (to facilitate data exchange)

Benefits of a Metadata Registry

- Provides the mechanisms for enabling global data acquisition and interchange, particularly across application areas. Data definitions and descriptions are precise to support reuse or multiple users of data.
- Documentation of data characteristics to support fully automated sharing of data, including locating, retrieving, and exchanging data.
- Provides uniform guidance for the identification, development, and description of elements and domains.
- Metadata registries provide universal means for organizing standard shareable elements thereby facilitating search, retrieval and optimal usage.

- Sets up common data standards between organizations. Exchange of data among organizations is facilitated with the common data standards.

Appendix C - Survey Results – Technologies Used

A wide variety of technology is used by the participating agencies. This variety is captured in the technology Table C-6 – Study 'Big Eight' Technology Count. When required the team would contact the data owner for data clarification. In some cases, as with the 202 "Unspecified", the data owners were not sure of the technology used.

Technology Type	Use Count
.NET	23
Active Server Pages	8
ActiveSync	1
Business Objects (DTO)	1
C	5
C++	13
COTS	27
DB Link (unspecified)	3
EDI	20
ETL	6
Filemaker Pro Server	10
FTP	75
IBM DB2	3
IBM LU 6.2	3
IBM LU2	2
IBM MQ	5
Java/XML	6
Mainframe[38]	328
Manual	156
MS Access	54
MS FoxPro	44
MS SQL Server	33
Multiple technologies	77
Natural/ADABAS	6
ODBC	3
Oracle	101
Peoplesoft Tools	13
Pro database	15
SAS	5
SMTP	1

[38] Mainframe represents a mix of technology responses that were so varied; it made better sense to lump together as one category. Examples of the responses are CICS, VSAM, and DB2.

Technology Type	Use Count
SWIFT	3
TCP/IP[39]	5
VB	15
VPN	2
Web services (unspecified)	47
XML (unspecified)	5
Unspecified	202

Table C-6 – Study 'Big Eight' Technology Count

[39] In the analysis it was unclear what was meant by TCP/IP since TCP/IP is a communication protocol. A couple of possibilities could have been FTP/SFTP over TCP/IP or a TCP/IP socket communication.

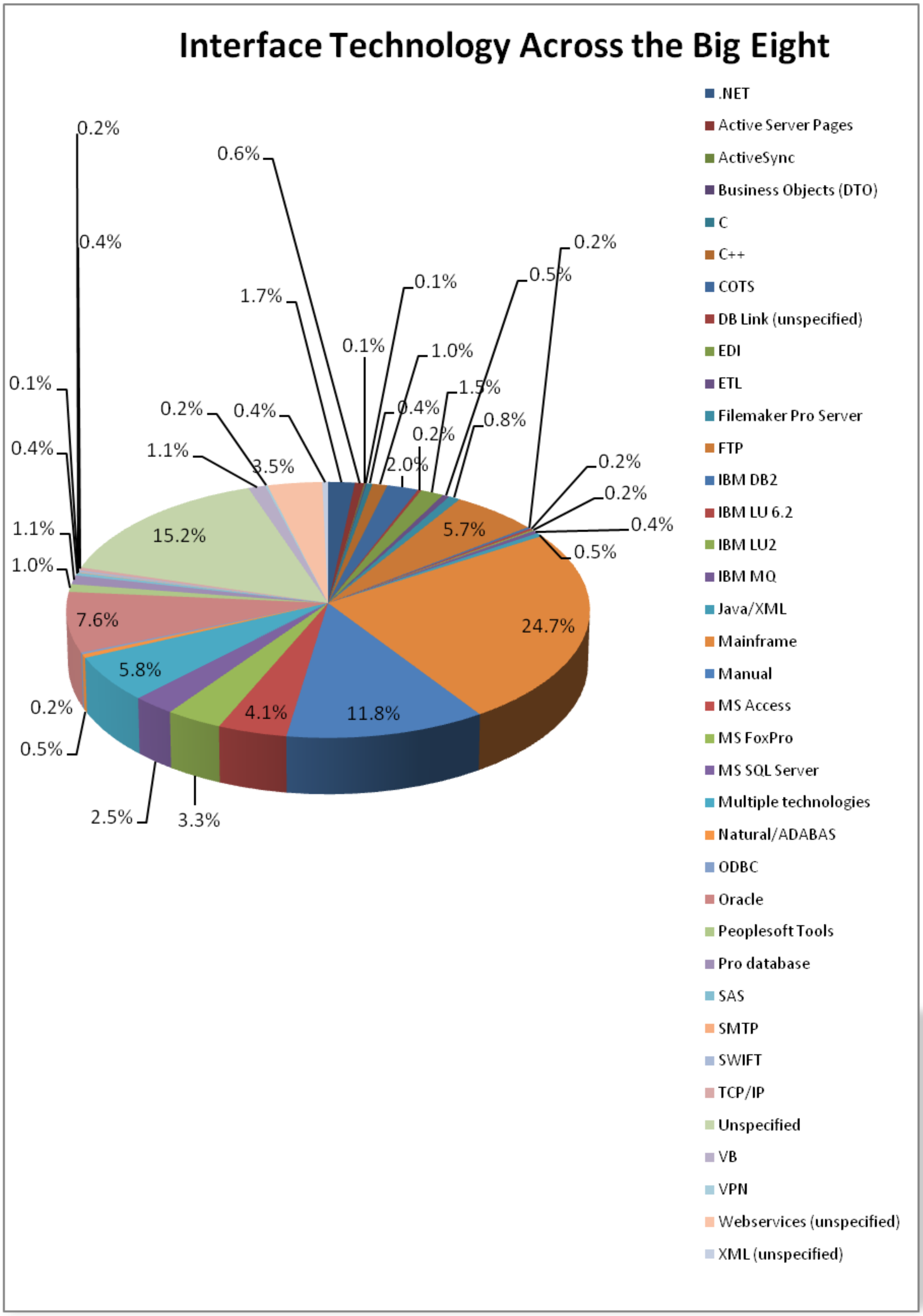


Chart C-1 – Agency Data Sharing Analysis

Technology by Agency

The following subsections present a breakdown by agency of all technologies employed by them. As one would expect, the bigger the agency, the more diverse set of technologies were employed.

Business, Transportation and Housing Agency

The Business, Transportation and Housing Agency (BTH), has the most diverse set of technologies. At the agency level, the IT infrastructure is decentralized. This is consistent with the fact that BTH has three very distinct businesses; Business, Transportation and Housing. As each department is considerably different, it is no surprise that the technologies used by each of these departments are quite different also.

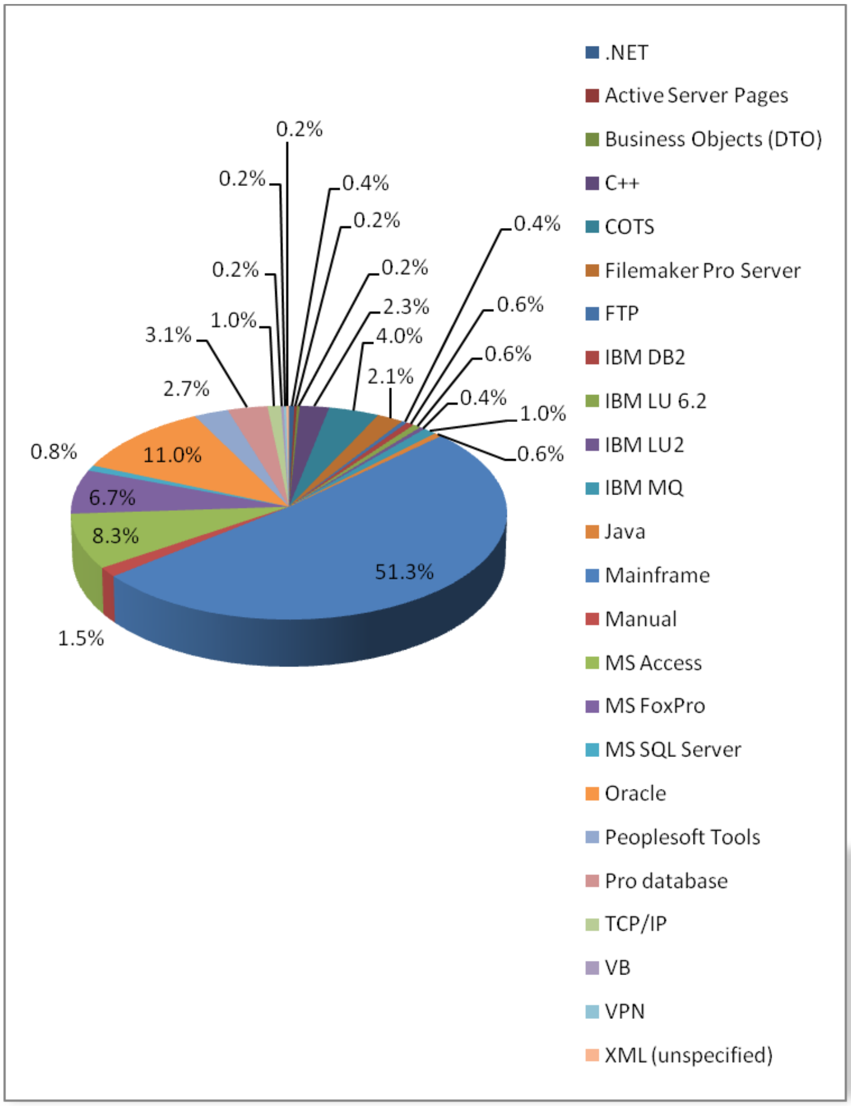


Chart C-2 – BTH Data Sharing Analysis

California Environmental Protection Agency
Initiatives to establishing data sharing across the departments are currently underway with the Unified Program System application. In addition they have adopted NIEM standard interface definitions for communication to the Federal Environmental Protection Agency. The technologies used for the interfaces that were reported by the California Environmental Protection Agency (CalEPA) were Java, Oracle or a combination of the two.

California Department of Corrections and Rehabilitation
The information that was provided by the California Department of Corrections and Rehabilitation (CDCR), is described in the chart below. The chart shows that they have a wide variety of technologies that they use in sharing their data. However, the vast majority of the information that is shared is between mainframe applications.

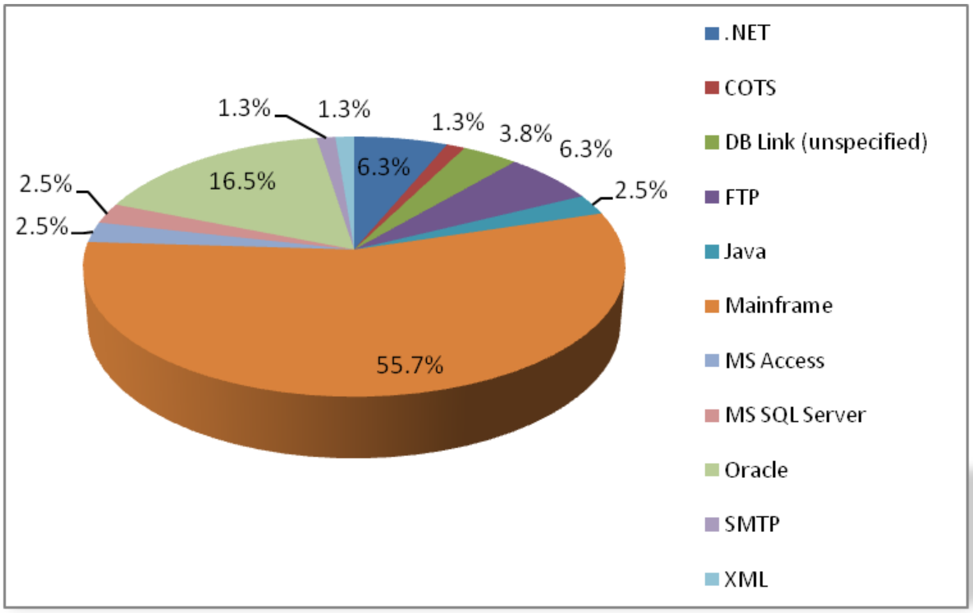


Chart C-3 – CDCR Data Sharing Analysis

California Department of Food and Agriculture
The California Department of Food and Agriculture (CDFA), provided data that indicates agency wide standardization. All interfaces that were reported leverage Microsoft's .NET technology. As IT standardization has been a focus of CDFA, their focus and their approach to standardize on technologies used to run their business, is an excellent cost saving move for them.

California Human and Health Services Agency
The California Health and Human Services Agency (CHHS), oversees twelve departments and one board that provide a range of health care services, social services, mental health services, alcohol and drug services, income assistance, and public health services to Californians from all walks of life. Even with the varying chartered services offered and significant number of people served CHHS has managed to limit the number of technologies used for data sharing. CHHS's strategic drive to move to web services is evidenced in their interface counts.

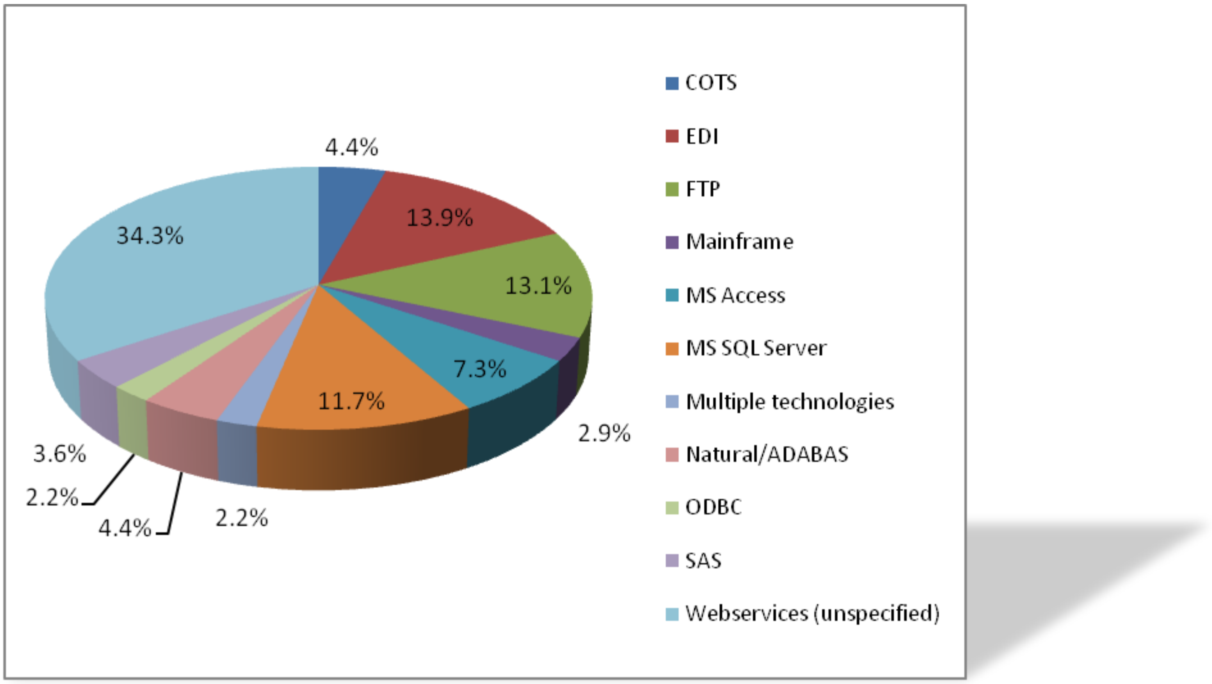


Chart C-4 – CHHS Data Sharing Analysis

California Natural Resource Agency
The California Natural Resource Agency (CNRA), is one of the smaller agencies surveyed for this data strategy. CNRA's size gives them the agility to drive toward technology standardization. CNRA appears to favor Microsoft technology. As seen earlier in the report, this is a great approach to drive down support costs when working with a hand full of technologies.

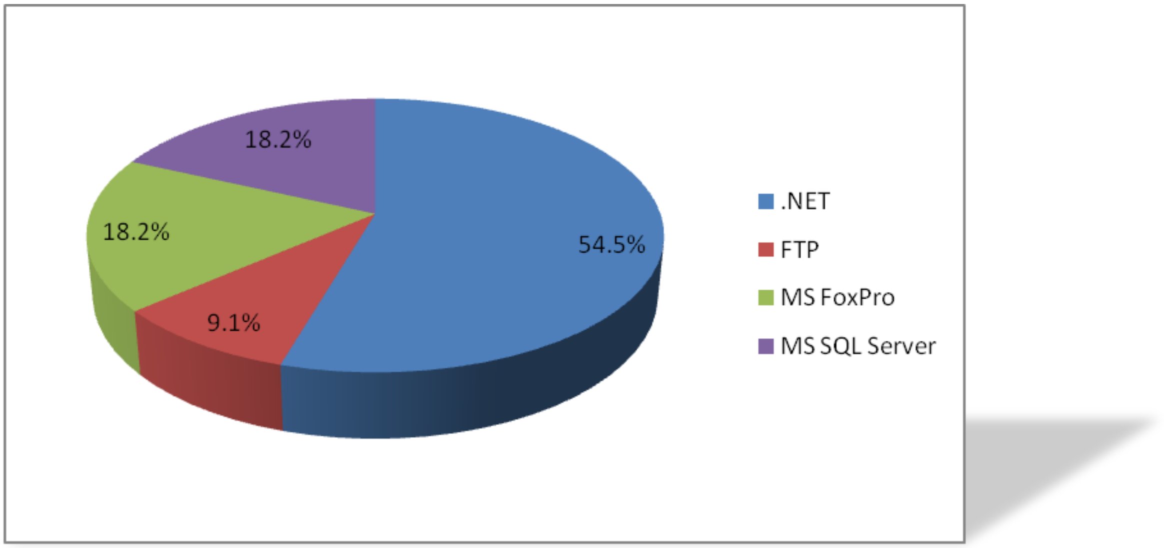


Chart C-5 – CNRA Data Sharing Analysis

Labor and Workforce Development Agency

The Labor and Workforce Development Agency (LWDA), is another one of the very large agencies. Their biggest department is Employment Development Department (EDD). The nature of EDD's business requires them to process large amounts of data and report that information to the Federal Department of Labor. Because of their federal ties, the primary technologies used are mainframe based, while COTS solutions and Oracle technology are being used to lesser degree. Transferring flat files via FTP is still prevalent within the agency.

Also, revealed in the research, were the many interfaces that required manual intervention. Examples of manual intervention would be keying in data by hand, cutting a CD-ROM or printing a file.

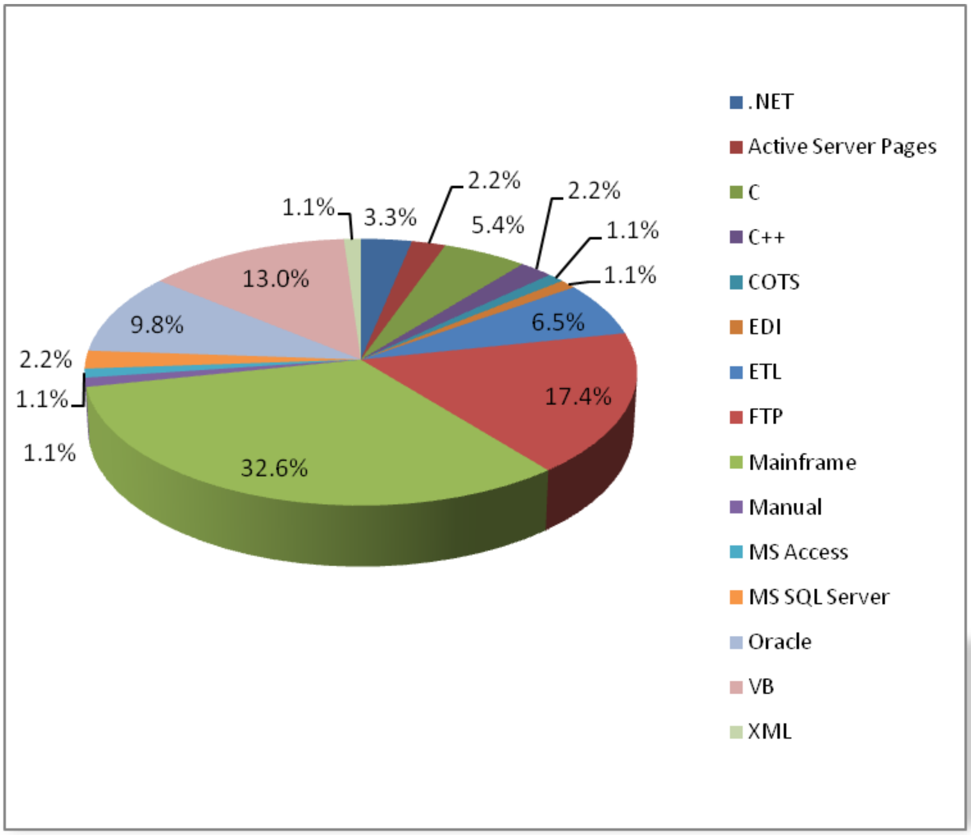


Chart C-6 – LWDA Data Sharing Analysis

State Consumer Services Agency

The State Consumer Services Agency (SCSA), is unique as an agency as its departments provide a wide variety of services. Since there is such diversity in the services provided by the agency the individual departments operate autonomously with respect to selecting and using software solutions. They have a wide variety of technologies employed to support their businesses. They too, have manual processes in which CD-ROMs are burned, magnetic tape is written or data is entered online.

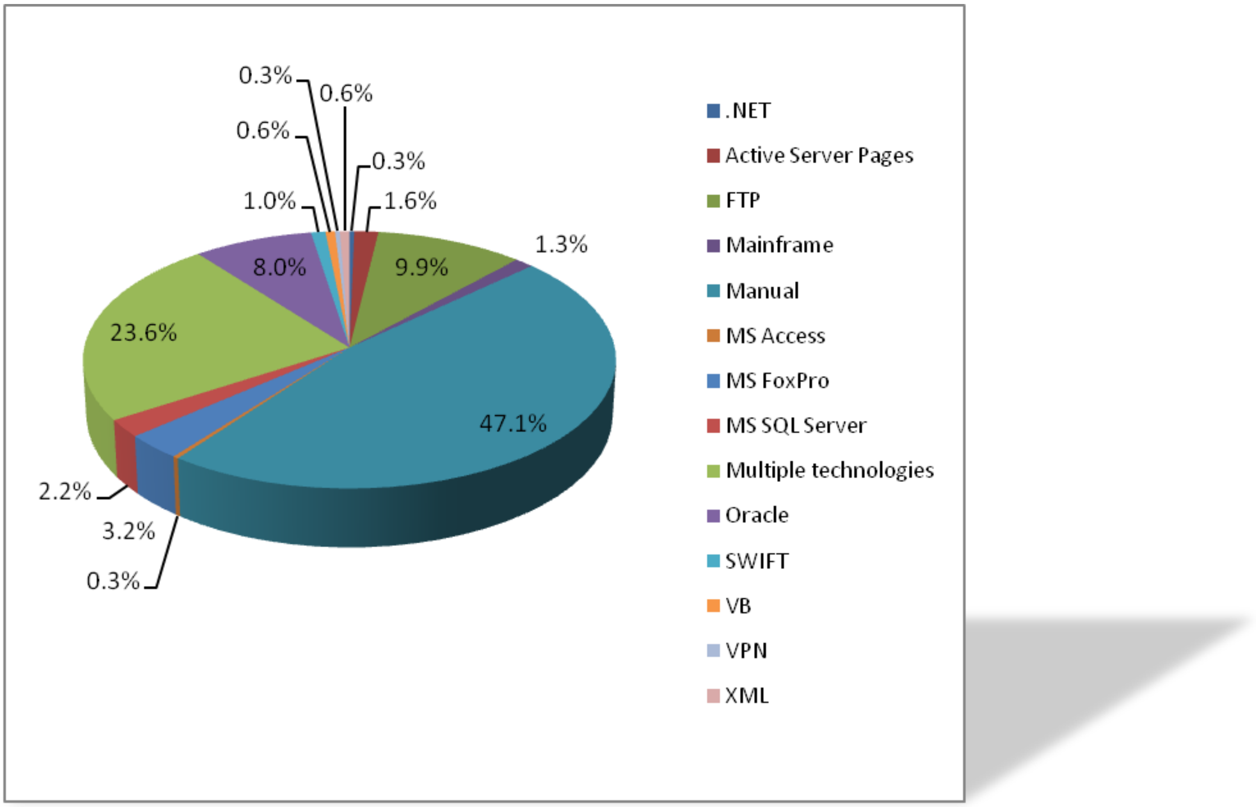


Chart C-7 – SCSA Data Sharing Analysis

This page intentionally left blank

Appendix D - DATA REFERENCE MODEL SPECIFICATION VERSION 2.0

The Data Reference Model specification is currently at version 2. Included is a copy of the specification here.

DRM_2_0_Final.pdf

This page intentionally left blank

Appendix E - Data Reference Model Extensible Markup Language Schema And Sample

DRM XML Schema

Draft_FEA_DRM_XML_Schema_v.0.3.pdf

The sample has been split into three sections representing the standardization areas of the DRM.

Data Description Section

DataDescription.pdf

Data Context Section

DataContext.pdf

Data Sharing Section

DataSharing.pdf

This page intentionally left blank

Appendix F - THIRD PARTY SCHEMA EVALUATION

Most software solutions provide data schemas for common data objects such as Facility, Location, Person, etc. More specialized solutions support data objects such as Law Enforcement, Medical Records, etc. Each solution below has a specialty and can be customized at least within their subject area(s).

National Information Exchange Model (NIEM) - http://www.niem.gov
NIEM is an XML based framework originally designed to support information sharing for justice, emergency and disaster management, public safety, and homeland security agencies. The standard is highly extensible and is easily adopted across a wider scope of agencies.

Global Justice XML Data Model (GJXDM) - http://www.it.ojp.gov/gjxdm
GJXDM is an XML based framework originating from the NIEM standard designed to support information sharing for criminal justice and public safety agencies.

Schools Interoperability Framework (SIF) - http://www.sifinfo.org
A data sharing framework for academic institutions from kindergarten through twelfth grade. SIF consists of an XML standard for defining educational data and a Service-Oriented Architecture (SOA) specification for exchanging data between institutions.

Sharable Content Object Reference Model (SCORM) - http://www.conform2scorm.com
A framework for web-based e-learning. SCORM was developed by the Advanced Distributed Learning (ADL) Initiative, which comes out of the Office of the United States Secretary of Defense.

Data Reference Model (DRM) - http://xml.coverpages.org/ni2005-12-28-a.html
DRM is an XML based framework part of the Federal Enterprise Architecture (FEA) designed to support information sharing within the United States federal government. DRM leverages several other standards such as ISO/IEC 11179.

Universal Data Element Framework (UDEF) - http://www.opengroup.org/udefinfo
A framework to support building of an enterprise wide controlled vocabulary.

Federal Enterprise Architecture (FEA) - http://www.whitehouse.gov/omb/e-gov/fea
FEA is a framework comprised of five separate standards designed to provide guidance for information technology efforts within the federal government.

- **Performance Reference Model (PRM)**
  A framework to measure performance and program contribution of IT investments
- **Business Reference Model (BRM)**
  A framework to describe day to day business operations
- **Service Component Reference Model (SRM)**
  A framework to help classify how well service components support business or performance objectives

- **Technical Reference Model (TRM)**
  A framework to help classify standards and technologies used to support and deliver service components
- **Data Reference Model (DRM)**
  A framework to describe data elements used within the interaction between the federal government and citizens

NIST Enterprise Architecture Model (NIST EA Model) - http://www.faa.gov/niac
A framework to organize, plan, and build business, information, and technology components. The architecture model has five overlapping layers:

- Business Architecture
- Information Architecture
- Systems Architecture
- Data Architecture
- Delivery Architecture

Treasury Enterprise Architecture Framework (TEAF) - http://www.ustreas.gov/offices/cio
A Zachman[40] based framework to support treasury business processes. The framework has a four layer view:

- Functional View (how, where, when)
- Information View (what, how much, how frequently)
- Organizational View (who, why)
- Infrastructure View (enabler)

Oracle Application Integration Architecture (AIA) - http://www.oracle.com/applications/oracle-application-integration-architecture.html
AIA offers packaged software for data, process and UI integration designed to provide an end-to-end solution. The AIA software components were designed to work together in a mix and match fashion, meaning they are easier to customize than typical standalone packaged products. AIA offers pre-built integrations (packaged solutions) as well as so called foundation packs (framework solutions).

Pre-built integrations are either direct integrations or process integrations packs:

- **Direct Integrations (DI)**
  Pre-built integrations that manage data flows and data synchronizations between specific applications
- **Process Integration Packs (PIPs)**
  Pre-built business processes across enterprise applications

[40] John Zachman developed a standard framework for enterprise architecture while working at IBM in the 1987.

Examples of cross application pre-built Process Integration Packs (PIPs):

- Oracle CRM to Siebel CRM
- Agile PLM to Oracle E-Business Suite
- Siebel CRM to Oracle Order Management (Order to Cash)
- Oracle CRM to Oracle E-Business Suite
- Demantra to Siebel CRM
- Siebel Life Sciences to Oracle Adverse Event Reporting System
- Siebel to Oracle Trade Promotion Management
- Siebel CRM to i-flex's FLEXCUBE Account Origination (Liability Products)

Advantages of AIA Foundation Packs:

- Simplify cross-application business process integrations by leveraging pre-built code designed for re-usability and configurability
- Code development is done within a standardized framework, and is leveraging existing Oracle and non-Oracle application investments
- Ability to utilize Service Oriented Architecture (SOA)

SAP NetWeaver Process Integration (SAP PI) - http://www.sap-xi.com
While SAP's product portfolio enables the company to provide turnkey solutions, the Process Integrator (PI) (formerly known as Exchange Infrastructure (XI) is a packaged product of their NetWeaver product group capable of exchanging enterprise application information between agencies.

Pervasive Software Data Integrator - http://www.pervasive.com/dataintegrator
Pervasive Software is a good example of an expanding packaged solution provider. Their Data Integrator product (formerly known as Data Junction) provides close to 250 data level integration solutions such as:

- File and Database Connectors
- Application and B2B Connectors
- Technology and Legacy Connectors

Vitria BusinessWare - http://www.vitria.com/BusinessWare
Vitria BusinessWare is a turnkey solution for order fulfillment, supply chain interactions, insurance claims processing, and other financial transactions. It utilizes a choreography approach to model, implement, monitor, and manage end-to-end processes spanning agencies and computer systems.

AIRS Standards for Professional Information & Referral - http://www.airs.org
A turnkey solution provided by the Alliance of Information and Referral Systems (AIRS) supporting the human services sector (e.g., 2-1-1 phone number as a partnership between AIRS and the United Way of America).

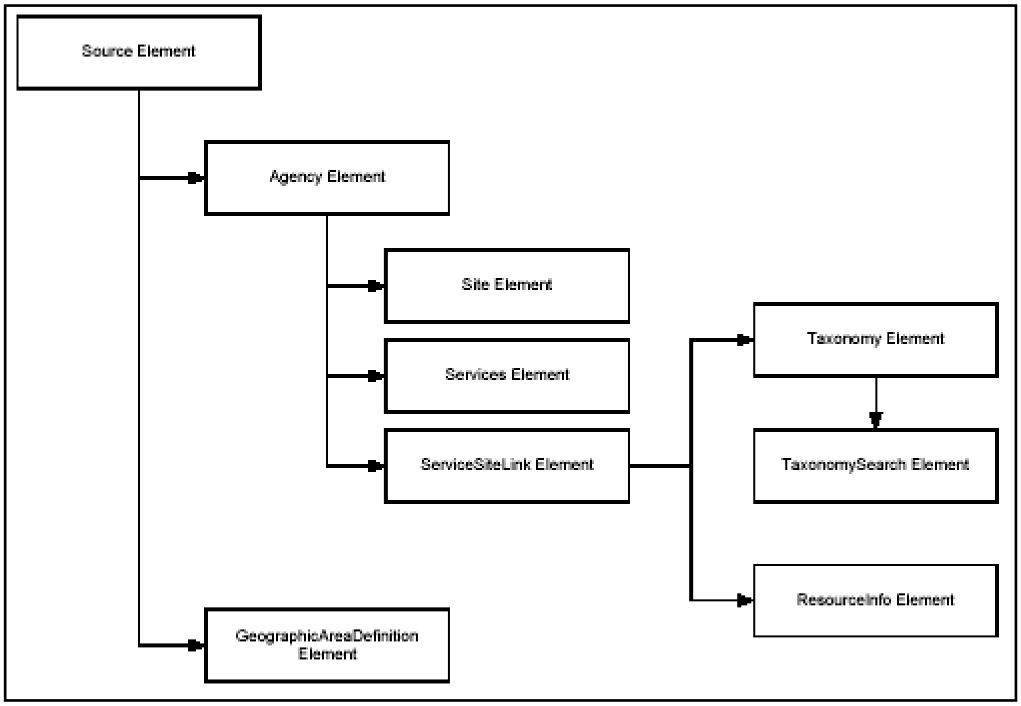


Figure F-1 - Simplified AIRS Schema

Appendix G - SAMPLE DATA SHARING MEMORANDUM OF UNDERSTANDING (MoU)

The objective of the document is to provide a template to aid in the development of data sharing agreements for agencies.

Guideline for
Establishing Data Exc

Sample DMV MoUs are provided below.

Supplemental
Addendum_Sample.d

Commercial AD-HOC
Agreement_122107.c

This page intentionally left blank

Appendix H - FEDERAL GEOGRAPHIC DATA COMMITTEE – STREET ADDRESS DATA STANDARD (WORKING DRAFT 2.0)

05-11.2ndDraft.CompleteDoc.pdf

This page intentionally left blank

Appendix I - GLOSSARY OF TERMS

Authentication – A process for verifying that a person or computer is who they say they are.

Authorization – Once authenticated it informs if a person or computer is permitted to a resource

CalBRM – Acronym for California Business Reference Model

CEAP – Acronym for California Enterprise Architecture Program

CDS – Acronym for California Data Services. The DaaS platform recommended in the strategy.

CIO – Acronym for Chief Information Officer

COI – Acronym for Community of Interest. In the context of State of California, when multiple departments from within one agency or from across multiple agencies come together to manage and maintain a common data asset then they form a Community of Interest.

CTPNS – Acronym for California Trading Partner Network Services. It enables state agencies interface with trading partners such as federal agencies.

DaaS – Acronym for Data-as-a-Service. A concept that emphasizes on using SOA to access data.

Data Context – Data context refers to any information that provides additional meaning to data. Data context typically specifies a designation or description of the application environment or discipline in which data is applied or from which it originates. It provides perspective, significance, and connotation to data, and is vital to the discovery, use and comprehension of data.

Data Dictionary - As defined in the *IBM Dictionary of Computing*, it is a "centralized repository of information about data such as meaning, relationships to other data, origin, usage, and format."

Data Element - A precise and concise phrase or sentence associated with a data element within a data dictionary (or metadata registry) that describes the meaning or semantics of a data element.

Data Governance - It refers to the operating discipline for managing data and information as a key enterprise asset.

Data Harmonization – It is the act of consolidating data from different sources according to the business rules that are established to enable a single, secure, validated, cleaned set of data

Data Integrity – It is an assurance for administrators and users that data is being accessed and modified only by authorized users

Data Management - It is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets.

Data Modeling – A structured method for representing and describing the data used in an automated system. Data modeling is often used in combination with two other structured methods, data flow analysis and functional decomposition, to define the high-level structure of business and information systems.

Data Ownership – It refers to both possession and responsibility for data.

Data Reference Model - The Data Reference Model (DRM) is a flexible and standards-based framework to enable information sharing and reuse across the federal government via the standard description and discovery of common data and the promotion of uniform data management practices. The DRM provides a standard means by which data may be described, categorized, and shared. These are reflected within each of the DRM's three standardization areas of Data Description, Data Context, and Data Sharing.

Data Warehouse – A central repository for significant parts of the data that an enterprise's various business systems collect specifically designed for reporting. It is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process, specifically providing data for Online Analytical Processing (OLAP) efforts.

DBA - Acronym for database administrator.

DR – Acronym for Disaster Recovery

DRM – Acronym for Data Reference Model. The DRM is a flexible and standards-based framework to enable information sharing and reuse across the federal government via the standard description and discovery of common data and the promotion of uniform data management practices. The DRM provides a standard means by which data may be described, categorized, and shared.

EA – Acronym for Enterprise Architecture

eDiscovery – It refers to a process in which data is sought, located, secured and searched with the intent of using it in a civil legal case.

ETL – Extract, Transform, and Load, which is a process to extract data from one source, transform (or cleanse) it, and load the result into another source. This is frequently part of populating a Data Warehouse.

FEA – Acronym for Federal Enterprise Architecture

FIPS - Federal Information Processing Standard (FIPS), one of many standards set by the Federal government for exchanging or processing data.

FTP – Acronym for File Transfer Protocol

GIS – Acronym for Geospatial Information Systems

HA – Acronym for High Availability

HIPA – Acronym for Health Information Privacy Act

HTTP – Acronym for Hypertext Transfer Protocol

ISO – Acronym for International Standards Organization

LoB – Acronym for Line of Business

Master data – It is harmonized structured data in the Shared data space.

MDM – Acronym for Master Data Management

Metadata – It is data about data.

Metadata Registry – It is a central location in an organization where metadata definitions are stored and maintained in a controlled method. Included in the registry are approved enterprise data definitions, representations (models, XML structures), links to physical constructs, values, exceptions, and data steward information.

MoU – Acronym for Memorandum of Understanding

NIST – Acronym for National Institute of Standards and Technology

OCIO – Acronym for Office of State CIO for the state of California

OLAP – Acronym for Online Analytical Processing. It is a reporting and data design approach intended to quickly answer analytical queries. Data to satisfy OLAP reporting and analysis needs are designed differently than data used for traditional operational use. Although OLAP can be achieved with standard relational databases, multidimensional data models are often used, allowing for complex analytical and ad-hoc queries with a rapid execution time.

OLTP – Acronym for Online Transaction Processing. It is a class of systems that facilitate and manage transaction-oriented applications.

RDBMS – Acronym for Relational Database Management System

Shareable Data – It is defined as data that is generated by one or more Line of Business (LoB) and is accessible by authorized users statewide

Shared data space – Is a secure network containing shared resources such as data assets and data services

SLA – Acronym for Service Level Agreement

SOA – Acronym for Service Oriented Architecture

SOI – Acronym for Service Oriented Integration

Structured data – It is data represented by a Data Model that provides explicit meaning to it. Example: Relational data in a RDBMS.

Security Assertion Markup Language (SAML) – Is an XML based security token that supports exchanges of authentication and authorization data between security domains.

Security Token Service (STS) – Is a service that is trusted by both the client and the Web service to provide interoperable security tokens. An example of a of security token is SAML.

Trading Partner – An organization with who the state of California has an ongoing business relationship. Example: Federal Agency.

UDDI – Acronym for Universal Description Discovery and Integration

Unstructured Data – It is data without any structure such as images, text.

WSDL – Acronym for Web Services Definition Language

WS-I – Acronym for Web Services Interoperability

XML – Acronym for Extensible Markup Language. It describes a class of data objects called XML documents and partially describes the behavior of computer programs which process them. XML is a subset of SGML, the Standard Generalized Markup Language. Among its uses XML is intended to meet the requirements of vendor-neutral data exchange, the processing of Web documents by intelligent clients, and certain metadata applications. XML is fully internationalized and is designed for the quickest possible client-side processing consistent with its primary purpose as an electronic publishing and data interchange format.

Appendix J - DATA GOVERNANCE PART II

NASCIO-DataGovernancePTII.pdf

This page intentionally left blank

Appendix K - Service Oriented Architecture Best Practices

NCOIF-soa-best.pdf

This page intentionally left blank

Appendix L - References

[i] "Federal Enterprise Architecture Program: The Data Reference Model" – http://www.whitehouse.gov/omb/asset.aspx?AssetId=561

[ii] "The ROI of SOA" - http://www.networkworld.com/techinsider/2005/101005-roi-of-soa.html

[iii] "Gartner Says Organizations Can Save More Than $500,000 Per Year by Rationalizing Data Integration Tools" - http://www.gartner.com/it/page.jsp?id=944512

[iv] "The Computerworld Honors Program: Department of Housing and Urban Development" - http://www.cwhonors.org/viewCaseStudy.asp?NominationID=273

[v] NIST Publication 800-100 Information Security Handbook, Chapter 6.

[vi] "The Privacy Act of 1974" – http://epic.org/privacy/1974act/

[vii] Full Text Search – Wikipedia - http://en.wikipedia.org/wiki/Full_text_search

[viii] Federal Enterprise Architecture Program: The Data Reference Model: Data Description Section - http://www.whitehouse.gov/omb/asset.aspx?AssetId=561

[ix] "OASIS: UDDI Specification" – http://www.uddi.org/pubs/uddi_v3.htm - Copyright © 2000 - 2002 by Accenture, Ariba, Inc., Commerce One, Inc. Fujitsu Limited, Hewlett-Packard Company, i2 Technologies, Inc., Intel Corporation, International Business Machines Corporation, Microsoft Corporation, Oracle Corporation, SAP AG, Sun Microsystems, Inc., and VeriSign, Inc. All Rights Reserved."

[x] "Web Services Description Language (WSDL) 1.1" – http://www.w3.org/TR/wsdl - http://www.w3.org/TR/2001/NOTE-wsdl-20010315 - Copyright© 2001 Ariba, International Business Machines Corporation, Microsoft

[xi] Steve Graham – The role of private UDDI nodes - IBM developerWorks. (http://www.ibm.com/developerworks/webservices/library/ws-rpu2.html)

[xii] Steve Graham – The role of private UDDI nodes in WebServices, Part 1: Six species of UDDI (http://xml.coverpages.org/grahamIBM-ws-rpu1.pdf)

[xiii] WS-Trust Specification - http://docs.oasis-open.org/ws-sx/ws-trust/200512/ws-trust-1.3-os.html - Copyright © OASIS® 1993–2007. All Rights Reserved. OASIS trademark, IPR and other policies apply.

[xiv] ColabWiki: Data Reference Model Version 2.11 2005 – Security and Privacy. http://colab.cim3.net/cgi-bin/wiki.pl?DataReferenceModel_Version2_11_2005/Security_And_Privacy#nid3X00

[xv] FGDC (2005) Street Address Data Standard (Working Draft 2.0) - http://www.fgdc.gov/standards/projects/FGDC-standards-projects/street-address/

[xvi] Federal Geographic Data Committee - FGDC-STD-011-2001 – US National Grid – http://www.fgdc.gov/usng/index_html/?searchterm=USNG

[xvii] B.W. Boehm, "A Spiral Model of Software Development and Enhancement," *IEEE Computer*, Vol. 21, No. 5, 1988, pp. 61-72.

[xviii] IBM, data Governance Website - http://www-01.ibm.com/software/tivoli/governance/servicemanagement/data-governance.html

[xix] “State Administrative Manual” – Chapter 5320.2 – RESPONSIBILITY OF OWNERS OF INFORMATION – http://sam.dgs.ca.gov/toc/5300/5320.2.htm

[xx] State Administrative Manual (SAM) Section 5100 – Electronic Data Processing (EDS) Policy - http://sam.dgs.ca.gov/TOC/4800/5100.htm

[xxi] Open Geospatial Consortium Inc. (2006) - OpenGIS® Web Map Server Implementation Specification

[xxii] Open Geospatial Consortium Inc. (2005) - Web Feature Service Implementation Specification